Simão Pedro Mendes Cruz Reis Paredes

Integration of Different Risk Assessment Tools to Improve the Event Risk Assessment in Cardiovascular Disease Patients

PhD thesis submitted to
Faculty of Sciences and Technology of University of Coimbra

Advisor:
Prof. Doutor Jorge Manuel Oliveira Henriques
(Assistant Professor of Faculty of Sciences and Technology of University of Coimbra)

Department of Informatics Engineering
Faculty of Sciences and Technology of University of Coimbra
May/2012

"The important thing is not to stop questioning. Curiosity has its own reason for existing."

Einstein, A., (1879–1955), German physicist

Acknowledgments

First, I would like to express my sincere gratitude to my advisor, *Prof. Doutor Jorge Manuel Oliveira Henriques*, for his support and guidance throughout this research. I have learnt from his experience as well as from his technical and scientific expertise. His scientific guidance and reviews were critical to the development of this work.

I also would like to thank *Prof. Doutor Paulo Fernando Pereira Carvalho* for his helpful comments as well as for making me feel effectively integrated in the HeartCycle FP7-216695 project.

I owe my deepest gratitude to my daughter *Carolina* and to my wife *Teresa*. They are crucial to my stability, without their strength this thesis would not have been possible.

I would like to thank my parents (*Elmano*, *Glória*), my brothers (*André*, *Filipa*) and my nieces (*Leonor*, *Mariana*, *Rita*) for all the support and nice relaxing moments.

I would like to thank the cardiologists *Dr. João Morais* from Leiria-Pombal Hospital Centre, Portugal; *Dr. Jorge Ferreira* from Santa Cruz Hospital, Lisbon/Portugal and *Dr. John Cleland* from Castle Hill Hospital, Hull/U.K for kindly providing the required data for the validation of the developed methodologies as well as for the valuable clinical thoughts.

I wish to thank the Coimbra Institute of Engineering/ Polytechnic Institute of Coimbra (ISEC/IPC); HeartCycle EU project (FP7-216695) and CISUC (Center for Informatics and Systems of University of Coimbra) for their support.

Finally, I would also like to show my gratitude to a group of people that contributed significantly for the success of this thesis: *Teresa Raquel, Francisco Pereira, Deolinda Rasteiro, Elizabeth Santos, Filipe Carvalho, Guilherme Figo, Leopoldina Almeida, Manuela Almeida, Pedro Bento, Margarida Pechincha, Paulo Matos Carvalho.*

Abstract

Each year cardiovascular disease (CVD) causes over 1.9 million deaths in the European Union (42% of all deaths), and contributes to health costs with a total estimated of €169 billion. These unaffordable social and health costs tend to increase as the European population ages. In this context, the correct prognosis of cardiovascular disease is a key factor to defeat the current statistics.

Some useful tools have been developed to predict the risk of occurrence of a cardiovascular disease event (e.g. hospitalization or death). However, these tools present some major drawbacks as they: *i*) ignore the information provided by other risk assessment tools that were previously developed; *ii*) consider (each individual tool) a limited number of risk factors; *iii*) have difficulty in coping with missing risk factors; *iv*) do not allow the incorporation of additional clinical knowledge; *v*) do not assure the clinical interpretability of the respective parameters; *vi*) impose a selection of a standard tool to be applied in the clinical practice; *vii*) may present some lack of performance.

This work aims to minimize the identified weaknesses, through the development of two different methodologies: *i*) combination of individual risk assessment tools; *ii*) personalization based on grouping of patients.

The former creates a flexible framework that is able to combine a set of distinct current risk assessment tools. The methodology is based on two main hypotheses: *i*) it is possible to implement a common representation (naïve Bayes classifier) of the individual risk assessment tools. Actually, current tools are diversely represented which does not facilitate their integration/combination. Moreover, these different representations are not suitable to deal with missing risk factors nor they can incorporate additional clinical knowledge; *ii*) it is possible to combine individual models exploiting the particular features of Bayesian probabilistic reasoning. The combination of individual models permits the creation of a global model that avoids the selection of a standard model as well as it can be adjusted to a specific population (optimized) through genetic algorithms operation.

The personalization based on the grouping of patients is proposed as an approach to enhance the performance of the risk prediction when compared to the one obtained with current risk assessment tools. This methodology is based on the evidence that risk assessment tools perform differently among different populations. Therefore, the main hypothesis that supports this methodology can be stated as: if the patients are properly grouped (clustered) it would be possible to find the best classifier for each patient.

Validation was performed based on three real patient datasets: *i*) Santa Cruz Hospital, Lisbon/Portugal, $N = 460$ ACS-NSTEMI patients; *ii*) Leiria-Pombal Hospital Centre, Portugal, $N = 99$ ACS-NSTEMI patients; *iii*) Castle Hill Hospital, Hull/U.K., $N = 426$ heart failure patients.

Considering the obtained results it is possible to state that the initial goals of this work were achieved, which makes it a valid contribution for the improvement of the risk assessment applied to cardiovascular diseases. However, other research directions should be pursued in order to improve the proposed methodologies and respective results.

Resumo

As doenças cardiovasculares provocam anualmente, aproximadamente 1.9 milhões de mortes na União Europeia contribuindo com um valor estimado de 169 mil milhões de euros para os custos de saúde. Estes custos associados às doenças cardiovasculares são insustentáveis e tendem a ser agravados dado o envelhecimento da população Europeia. Neste contexto, o prognóstico das doenças cardiovasculares é um factor chave para inverter as actuais estatísticas.

Existem algumas ferramentas muito úteis que foram desenvolvidas com o objectivo de avaliar o risco de ocorrência de um evento (hospitalização ou morte) originado por doença cardiovascular. No entanto, apresentam algumas lacunas importantes uma vez que: *i*) ignoram a informação disponibilizada por outras ferramentas de avaliação de risco previamente desenvolvidas; *ii*) individualmente consideram um número limitado de factores de risco; *iii*) têm dificuldade em lidar com factores de risco em falta; *iv*) não permitem a incorporação de conhecimento clínico adicional; *v*) podem não ser clinicamente interpretáveis; *vi*) requerem a selecção de uma ferramenta para aplicação na prática clínica; *vii*) apresentam alguns problemas de desempenho na predição do risco.

Este trabalho pretende contribuir para reduzir as fragilidades identificadas, através do desenvolvimento de duas metodologias: *i*) combinação de ferramentas de predição de risco; *ii*) personalização baseada no agrupamento de pacientes.

A primeira metodologia permite criar um modelo global tendo por base a combinação de ferramentas de avaliação de risco. Esta abordagem é baseada essencialmente em duas hipóteses: *i*) é possível implementar uma representação comum (classificador naïve Bayes) das ferramentas de avaliação de risco. Com efeito, as ferramentas disponíveis são representadas de forma diversa o que não facilita a sua integração/combinação. A dificuldade destas ferramentas em lidar com valores em falta assim como a sua incapacidade de incorporar conhecimento clínico adicional são factores adicionais que justificam a criação de uma representação comum; *ii*) é possível efectuar a combinação dos modelos individuais com base nas características específicas da inferência de Bayes. Deste modo, a combinação de modelos individuais

permite a criação de um modelo global que não só evita a necessidade de seleccionar uma ferramenta para utilização na prática clínica como também permite o ajustamento a uma população específica (optimização) através da operação de algoritmos genéticos.

A personalização da predição do risco com base no agrupamento de pacientes é proposta como uma abordagem para melhorar a avaliação do risco. Esta metodologia é baseada na evidência de que o desempenho das ferramentas de avaliação de risco varia em função das características específicas da população. Assim, a hipótese que suporta esta metodologia pode ser enunciada da seguinte forma: se os pacientes forem devidamente agrupados então é possível encontrar o classificador mais adequado a cada paciente.

A validação foi efectuada com base em três conjuntos de dados reais: *i*) Hospital de Santa Cruz, Lisboa/Portugal, $N = 460$ pacientes ACS-NSTEMI; *ii*) Centro Hospitalar de Leiria-Pombal, Portugal, $N = 99$ pacientes ACS-NSTEMI; *iii*) Castle Hill Hospital, Hull/U.K., $N = 426$ pacientes com insuficiência cardíaca.

Com base nos resultados obtidos, é possível afirmar que os objectivos iniciais deste trabalho foram atingidos, o que demonstra que esta tese é uma contribuição válida para a melhoria dos sistemas de avaliação de risco aplicado à doença cardiovascular. No entanto, devem ser exploradas outras linhas de investigação de forma a melhorar a metodologia proposta e consequentemente os resultados obtidos.

Table of Contents

Abbreviations and Notation

Abbreviations

ACS	Acute Coronary Syndrome
CAD	Coronary Artery Disease
CHD	Coronary Heart Disease
CPT	Conditional Probability Table
CVD	Cardiovascular Disease
DAG	Directed Acyclic Graph
DT	Decision Tree
EHR	Electronic Health Records
EU	European Union
HF	Heart Failure
GA	Genetic Algorithms
LPHC	Leiria-Pombal Hospital Centre
MAR	Missing at Random
MCAR	Missing Completely at Random
NMAR	Not Missing at Random
NSTEMI	Non-ST Segment elevation
PDA	Personal Digital Assistant
SVM	Support Vector Machines

Notation

p	Number of risk factors
$\mathbf{x}$	Instance (vector of p risk factor -value pairs)
$\overline{\mathbf{x}}$	Vector of the mean values of risk factors
x	Individual Risk factor/attribute
Υ	Set of instances
ψ	Set of risk factors (attributes)

c	Output class
C	Set of output classes
m	Number of mutually exclusive classes
D	Set of labeled instances $(\mathbf{x}, c)$
N	Number of instances
R	Set of real numbers
R^p	Set of real *p–dimensional* vectors

Risk assessment tools' derivation

$S(t)$	Survival function
t	Time
T	Survival time
$\boldsymbol{\beta}$	Vector of the proportional hazard regression coefficients
$h(t)$	Hazard rate function
$h_0(t)$	Baseline hazard function

Logic based algorithms

d_i	Decision node i
b_{ij}	Branch from decision node i to decision node/leaf j
L_i	Leaf i
q_i	Decision node disjoint outcomes of decision node d_i
v	Individual split

Perceptron based techniques

b	Neuron model's bias
g	Neuron model's activation function
h	Neuron model's combination function
u	Neuron model combination function's output
y	Neuron model's output
$\mathbf{w}$	Weights vector
z	Number of hidden layers
L	Neural network Layer

Instance Learning

$d(\mathbf{u}, \mathbf{v})$	Distance between data vectors $\mathbf{u}, \mathbf{v}$

Support vector machines

H	Support vector machine hyperplane
d_+	Distance to the closest positive instance
d_-	Distance to the closest negative instance
$\boldsymbol{w}$	Vector normal to the hyperplane
ξ	Margin
ζ	Misclassification penalty

Model output combination

$A(\cdot)$	Learning algorithm
M	Classifier/model
J	Training data set
O	Testing data set
$P(c \mid M_i)$	Probability of class c exclusively based on model M_i
$P(M_i \mid J)$	Probability of model M_i being correct given the training dataset J
$P(J \mid M_i)$	Likelihood of model M_i to generate J

Optimization

$f(x)$	Objective function
l_{ri}	Equality constraints
l_{rj}	Inequality constraints
f_i	Fitness of individual i
$P_{sel}(i)$	Probability of an individual i to be selected
$\boldsymbol{b}$	Vector of ordered probabilities
μ	Population's size
$\mathbf{p}^i$	Parent i
$\mathbf{ch}^i$	Child i

Validation

σ	Population's standard deviation
s	Sample's standard deviation
se	Standard error
B	Number of bootstrap samples
$\hat{\theta}$	Estimator of θ

Missing information

$\overline{W}$	Natural variability of data
W_i	Variance of individual estimate
n	Number of imputed data sets
U	Uncertainty due to imputation

Dimensionality Reduction

$\mathbf{y}$	Instance (lower dimensional (q) representation)
$\boldsymbol{X}_{p\times N}$	Matrix of the N original instances
$\mathbf{Y}_{q\times N}$	Matrix of the N lower dimensional instances
$\boldsymbol{W}_{q\times p}$	Matrix that contains the q weight vectors of dimension p
$\boldsymbol{\Sigma}_{p\times p}$	Covariance matrix of $\boldsymbol{X}_{p\times N}$
λ_i	Eigenvalue i
$\boldsymbol{\Lambda}_{p\times p}$	Diagonal matrix of the ordered eigenvalues $\lambda_1 \leq ... \leq \lambda_p$
$\mathbf{U}_{p\times p}$	Orthogonal matrix that contains the eigenvectors

Clustering

Q_j	Objective function (k-means algorithm)
K	Number of clusters
$\mathbf{c}_j$	Center of cluster j
G_j	Cluster j
χ	Neighborhood (density attenuation)
ρ_i	Density of instance $\mathbf{x}_i$
S	Subsample of input data

u_{ji}	Element of membership function (k-means algorithm)
Z	Number of partitions

Probabilistic Reasoning/Bayesian Classifiers

X, Y	Random variables
$\mathbf{X}$	Vector of random variables that contains the p observed attributes
V	DAG nodes
E	DAG edges
$\mathcal{G}$	DAG
$\theta_{\mathcal{G}}$	Parameters of $\mathcal{G}$
Pa_i	Parents of node X_i
θ_{ij}	$P(X_i \mid Pa_i = j)$
θ_{ijk}	$P(X_i = k \mid Pa_i = j)$
N_{ijk}	Number of cases in the training data such that $X_i = k$ and $Pa_i = j$
I_f	Interval frequency
I_n	Interval number
Γ	Sum of individual models' weights
ϑ	Sum of individual models' weights that contain X_i
β_N	Neighborhood (optimization)
δ_{kr}	Variation on category k of attribute X_i given the class r
M_i	Individual model i
S_{M_i}	Subset of variables that belong to M_i
$M_{i,s}$	Model i subsample s

List of Figures

List of Tables

1. Introduction

Cardiovascular disease (CVD) is caused by disorders of the heart and blood vessels, including coronary heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. This disease is the world's largest killer, responsible for 17.1 million deaths per year *(WHO, 2009)*.

In fact, each year cardiovascular disease (CVD) causes over 1.9 million deaths in the European Union (42% of all deaths), and contributes to health costs of €105 billion (with a total estimated cost of €169 billion to the EU economy). Coronary heart disease (CHD), approximately half of all CVD deaths, is the single most common cause of death in Europe, and individually results in direct health costs of €23 billion. These costs include treatment of conditions and events resulting from CHD, which include myocardial infarction and heart failure (HF). There are about 10 million patients treated for heart failure (often resulting from CHD) in the EU, resulting in 2% of the total health care costs. Currently heart failure is the most frequent cause of hospitalization among individuals over 65 originating in hospitalization costs that are significantly higher than those of cancer and myocardial infarctions combined *(EHN, 2008)*.

Furthermore, the population of the EU and the western world is aging. The number of elderly people aged 65-79 will increase approximately by 37% by 2030 *(CEC/EU, 2005)*. Thus, it is recognized that this demographic change in the population will result in unaffordable health costs.

In this context the current health care paradigm must be changed. In fact, the health system has to move from reactive care towards preventive care and simultaneously transfer the care from the hospital to patient's home. According to European Heart Network around 80% of CHD are preventable *(EHN, 2009)*, which illustrates that the improvement of preventive health care can originate important benefits and reduce the incidence of cardiovascular diseases.

Health telemonitoring systems are assuming a critical importance in improving the preventive health care. They allow the remote monitoring of patients who are in different locations away from the health care provider. A set of devices installed in the patient's house (mainly interfaces and sensors) can be very valuable for the management of the patient condition. Clinical data (weight, blood pressure, electrocardiogram, etc.) can be collected, processed or sent to the care provider. As a result of the data processing, feedback can be provided directly to the patient as well as to the care provider, which may include the generation of alarms. Computational interfaces (PDA, Smartphone, etc.) may be used to obtain some additional subjective information from the patient as well as to provide feedback to patients.

This remote monitoring is more challenging to the care provider, as the reliability/quality of the clinical decision must be guaranteed in order to optimize therapy. At the same time, the patient has a crucial importance in this communication/decision process and in that perspective becomes more responsible for his/her health. Figure 1.1 presents one example of these systems (HeartCycle Project):

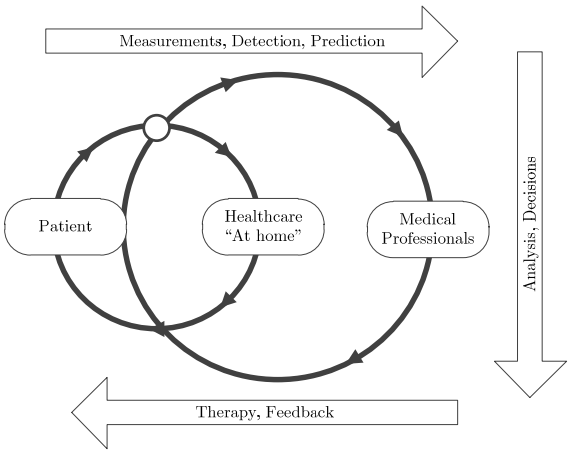


Figure 1.1- Patient and professional loop – HeartCycle project (Reiter, 2009).

HeartCycle project provides a closed-loop disease management solution being able to serve both Heart Failure (HF) patients and Coronary Artery Disease (CAD) patients, including possible co-morbidities, hypertension, diabetes and arrhythmias. This is achieved by multi-parametric monitoring, analysis of vital signs and other measurements *(Reiter, 2009)* .

The system contains: *i*) a patient loop interacting directly with the patient to support his daily treatment. It shows the patient's health development including

treatment adherence and respective effectiveness; *ii*) a professional loop involving medical professionals alerting them to the need of revisiting the patient's care plan and possible adverse events. The professional loop connects the patient loop system with the hospital information system, in order to ensure optimal and personalized patient care.

In both loops the risk assessment of and event occurrence (death, myocardial infarction, hospitalization, disease development, etc.) due to cardiovascular diseases is a critical issue. In fact, the correct prognosis[1] of cardiovascular disease is a key factor to help clinical professionals to identify the best treatment to each patient as well as to motivate the patient increasing the treatment compliance with the corresponding health benefits.

Several risk assessment tools[2] were developed to assess the probability of occurrence of a CVD event within a certain period of time (months/years). According to the risk assessment tool, two types of risk may be calculated: absolute risk, i.e., probability of developing a CVD event over a given period of time (e.g. 10 years), and a relative risk, i.e., risk of someone developing a CVD event that has risk factors compared to an individual of the same age and sex who does not, during a certain period of time *(NVDPA, 2009).*

Additionally, available risk assessment tools differ on the assessed period of time (months/years), predicted events (death/non-fatal), disease (coronary artery disease, heart failure, etc.), risk factors considered in the model, patient's conditions (ambulatory patients, hospitalized patients, cardiac transplant candidates, etc.).

These tools are derived from clinical datasets, usually through statistical methods that require a long monitoring period of the population sample[3]. Then, a predictive model is derived and is applied to classify new instances[4].

---

[1] Prognosis relates to the probability or risk of an individual developing a particular outcome over a specific time. This assessment is based on both clinical / non-clinical information that is available at the time of the prediction (Moons, 2009).

[2] In order to clarify, risk assessment models that have been statistically validated and are available in literature are going to be designated through this work as **risk assessment tools**.

[3] Extraction of the respective survival function, which captures the probability $P(\bullet)$ that an individual survives beyond a specified time $S(t) = P(T > t)$, t: specified time; T : time of event.

[4] An instance is described by a list of features which is the designation of a variable/value pair. A variable (attribute) is a quantity that describes a particular aspect of an object of the world. A dataset is a collection of instances *(Visweswaran, 2007)*

As mentioned, these risk assessment tools are very important to adapt the patient's personal care plan according to a given specific risk-reduction effort *(Bertrand, 2002) (Graham, 2007) (Koopman, 2008).*

Actually, this risk assessment has a positive impact on the management of an individual patient, since it contributes to close monitoring of the patient based on consistent data. In this way, it may be easier for the medical professional to adapt the personal care plan, according to a specific risk-reduction effort, as well as, tailoring the frequency of clinical follow-up visits.

In addition, this assessment also contributes to help medical professionals in managing the patient population. They have more information to identify those patients that need urgent hospitalization, those that need urgent review of respective care plans (lack of treatment, over treatment situations…) and those that correspond with the expected condition.

The feedback to the patients may also be a key factor in the patient's motivation, which could lead to an increase of treatment compliance since patients understand that their actions are crucial in modifying their own health status. Therefore, patients can learn about their personal risk assessment as well as the lifestyle changes they should make to reduce their cardiovascular event risk.

1.1 Motivation

In spite of the importance of these risk assessment tools they present important weaknesses that must be circumvented in order to improve the risk prediction.

In effect, current tools are usually developed without considering the information provided by other risk assessment tools that were previously developed.

Another important limitation of the current risk assessment tools is related with the limited number of risk factors that each tool considers individually. The evolution of Electronic Health Records (EHR) contributed to the availability of a large set of data from the patients. As a result, more accurate and more accessible data allows the improvement of medical diagnosis *(Dreiseitl, 2005)*. Consequently, it is not reasonable that CVD risk assessment tools do not take advantage of the available information.

Typically, one of the current risk assessment tools should be selected as a standard tool to be applied in the daily clinical practice. This selection may be very

difficult as the performances of these tools may vary according to the characteristics of the specific population. Additionally, the technical opinion of each cardiologist is also an important aspect in this selection process.

These tools have difficulties in coping with incomplete information (missing risk factors). This is a very important limitation, since the occurrence of missing risk factors is a very frequent problem in health records. According to Khanna *(Khanna, 2005)*, *"… information on patients such as demographic data, medical history, treatments, test results, and family structure is often unavailable when a doctor greatly needs".*

CVD risk assessment tools do not allow the incorporation of empirical clinical knowledge. This is another flaw that must be defeated. Physicians should have the possibility to incorporate direct clinical knowledge into the prediction model in addition to the information provided by the considered risk factors.

Moreover, these tools often present some performance limitations. They are typically developed for an average patient, which results in a lack of personalization. In fact, current risk assessment tools frequently present sensitivity/specificity[5] values that do not assure a proper classification of a specific patient.

The inability to capture the dynamics of CVD risk evolution is also recurrently recognized as a glitch of current risk assessment tools *(Visweswaran, 2007),* as they cannot capture the risk evolution that results from the changes on some of the considered risk factors.

The main motivation of this thesis is to reduce some of the identified weaknesses of the current CVD risk assessment tools, namely:

- To consider the available knowledge. Rather than to derive a new model, the proposed approach aims to combine current CVD risk assessment tools;
- To avoid the need to choose a risk assessment tool as a standard tool, the combination allows the selection of one or more tools to make the risk assessment;
- To allow the consideration of a higher number of risk factors;
- To cope with missing information (missing risk factors);
- To incorporate empirical clinical knowledge (new risk factors) that physicians decide should be ideally integrated;

---

[5] Sensitivity $SE = TP / (TP + FN)$; Specificity $SP = TN / (TN + FP)$; TP: True Positive; TN: True Negative; FN: False Negative; FP: False Positive.

- To assure the clinical interpretability of the model;
- To improve the performance of the risk assessment when comparing it to the one achieved by the individual current risk assessment tools.

The inability to capture dynamics of the risk evolution is not explored in this work. However, it is a very important issue that should be developed in future research.

1.2 Overview

Two main methodologies are proposed for achieving the referred goals: *i*) combination of individual risk assessment tools; *ii*) personalization based on grouping of patients.

Combination of Individual Risk Assessment Tools

This approach aims to combine individuals risk assessment tools and it is based on two main hypotheses:

- It is possible to create a common representation of individual risk assessment tools. Current risk assessment tools are diversely represented (charts, equations, scores, etc.) which does not facilitate their integration/combination. Additionally, these kinds of representations are not suitable to deal with missing risk factors nor can they incorporate additional clinical knowledge. Therefore a common representation must be simple in order to easily allow the integration of the different individual models[6] and should have the required flexibility to incorporate additional variables. Moreover, its parameters/rules must be clinically interpretable;
- It is possible to combine individual models, which is the main focus of this thesis. The ability of combining available knowledge from various sources is useful since it creates a flexible framework. The combination of individual models also permits the implementation of optimization methodologies to increase the CVD risk prediction performance. Thus, the clinician may take advantage of this overall knowledge.

[6] Individual models are the representations of individual risk assessment tools.

In this context, a methodology is developed and it can be briefly described, as presented in Figure 1.2:

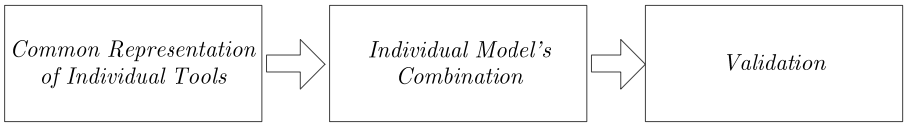


Figure 1.2 – Combination of individual risk assessment tools methodology.

The first step of this approach is the selection of current available risk assessment tools that are relevant in the CVD risk context. A common representation based on a machine learning classification algorithms[7] must be created in order to be applied to all the individual tools. The classifier must be selected considering not only that the individual models have to be combined but also have to deal with missing risk factors. Moreover, this common representation must assure the clinical interpretability of the model.

Individual models' combination is the essential step of the proposed approach. Rather than to derive a new global model, the goal is to create a model that can incorporate information from individual systems and/or directly from the physician. The global model that results from the combination scheme must be derived regarding the available input risk factors and the individual models' selection criteria. For instance, if one individual model does not have any of its input values available, then that model should not be considered for integration in the combination scheme. This innovative approach allows a very flexible model which is able to incorporate a variable number of input risk factors, it joins empirical clinical knowledge and it avoids the necessity of choosing a particular model as a standard model for the clinical practice. However, the clinical relevance of a CVD risk prediction system depends directly of its performance. Optimization techniques[8] are adopted in this phase to increase the global model's performance (maximize sensitivity and maximize specificity).

The third phase is validation that is determinant to evaluate the potential clinical importance of the proposed methodology. This phase is performed based on real data and it intends to be as inclusive as possible.

---

[7] Algorithms that learn how to assign the correct output's class label to testing instances. These algorithms can be based on neural networks, decision trees, Bayesian classifiers, support vector machines, k-nearest neighbor.

[8] Genetic algorithms.

Personalization Based on Grouping of Patients

This approach addresses, through a grouping strategy, the problem of the low performance exhibited by the current risk assessment tools.

The methodology is based on the evidence that risk assessment tools perform differently among different populations. The variation of performance indicates that a specific risk assessment tool may have a good performance within a given group of patients and performs poorly within another group. Thus, the main hypothesis that supports this methodology can be stated as: if the patients are properly grouped (clustered) it would be possible to find the best classifier for each group. Figure 1.3 presents the developed methodology:

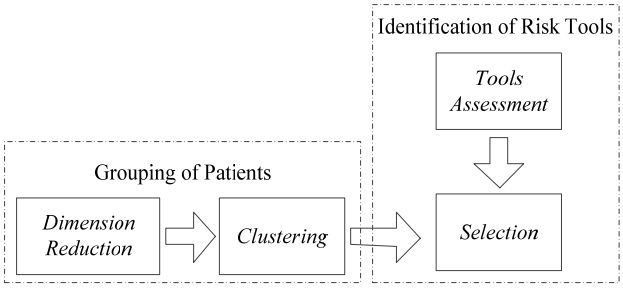


Figure 1.3 – Personalization based on grouping of patient's methodology.

The first phase is responsible for the proper grouping of patients and comprises two steps: *i*) dimensionality reduction procedure that is applied to reduce the number of variables required to the characterization of each patient; *ii*) clustering which is responsible for the creation of the patients' groups based on the information provided by the previous procedure. The second phase, attempts to select the most appropriate current risk assessment tool for each group of patients such that the CVD risk of a patient that belongs to a given group can be accurately estimated.

1.3 Contributions

The scientific goals of this thesis are the research, development and clinical validation of innovative models for improving the CVD event risk assessment. The main contributions can be detailed according to the two developed methodologies.

Combination of Individual Risk Assessment Tools

All the selected CVD risk assessment tools that are currently described through different methods (equations, scores, etc.) were represented based on Bayesian classifiers. These classifiers[9] have been chosen mainly due to their simplicity, to their capability of coping with missing information as well as to the possibility of incorporating empirical clinical knowledge.

The combination of individual models was performed through the fusion of the individual models' parameters. The resultant parameters were adjusted based on an optimization procedure that was carried out with genetic algorithms (GA).

Three data sets from three different hospitals were used for validation purposes in patients with different conditions: *i*) Santa Cruz Hospital – Lisbon, Portugal (460 patients); *ii*) Leiria-Pombal Hospital Centre, Portugal (99 patients); *iii*) Castle Hill Hospital – Hull, UK (426 patients). The methodology was validated to Coronary Artery Disease (CAD) also designated by Coronary Heart Disease (CHD) patients and Heart Failure (HF) patients.

Several new contributions of this dissertation to this research area can be identified:

- Creation of a methodology that considers the available knowledge provided by the current risk assessment tools. This is an important contribution which avoids the discarding of the available information originated by the previously developed tools;
- Development of an original combination strategy that takes into account the risk factors that belong to the different individual risk assessment tools. This has several implications, since the global model that results from the combination:
  - Allows the consideration of a higher number of risk factors. In fact, the number of risk factors depends directly on the individual models selected for the combination scheme;

[9] Bayesian classifiers belong to the category of Probabilistic/Statistical learning algorithms since they implement a probability model, which provides a probability that an instance belongs to an output class rather than a deterministic classification.

- Avoids the choice of a "standard model". The definition of a standard model used in the clinical practice can be very difficult, since there might not be a consensus about the model applicable. The combination approach is very interesting because more than one model can be used simultaneously to predict CVD risk;

- Ability to deal with missing information (missing risk factors). Probabilistic reasoning, that is the basis of the Bayesian inference mechanism, is well adapted to deal with missing information;
- Incorporation of empirical clinical knowledge. The developed combination scheme allows the combination of individual models, statistically derived and/or directly defined by the physician, e.g. influence of a specific risk factor not covered by current CVD risk models.

These methodological improvements of the CVD risk assessment systems have direct consequences in the clinical practice. Two different scenarios of this assessment use case can be identified:

- During ambulatory care/inpatient care, the cardiologist assesses the risk of a CVD event of a specific patient. The cardiologist does not have chance of validating[10] the performed prognosis (no available dataset of that specific population or model that fits that population);
- During ambulatory care/inpatient care, the cardiologist assesses the risk of a CVD event of a specific patient. The cardiologist has the possibility of validating the performed prognosis (available dataset of that specific population or model that fits that population).

1. Validation is not possible

In this case the current assessment procedure can be described as follows:

- The cardiologist assesses the risk of an event based on one or more of the current risk assessment tools. If there are several tools involved, a combination scheme must be implemented, usually a voting strategy. This assessment has some limitations, as the physician cannot incorporate empirical clinical knowledge into the tools, he has to replace missing values

[10] The correct validation of the risk assessment requires a proper set of data. Alternatively, the validation may be performed based on the empirical knowledge of the cardiologist.

by a specific value and he is restricted to the risk factors that are considered by the individual tools.

The proposed strategy in this thesis is more flexible as it allows the selection of different individual tools for the prediction. Besides this ability of considering a higher number of risk factors, the potential difficulty of choosing a tool to adopt in the clinical practice is also eliminated. The Bayesian inference mechanism does not require any imputation for the missing value. The ability of incorporating empirical clinical knowledge must also be stressed. These are important advantages of the proposed methodology in both scenarios. However, the unavailability of data or a model that fits that specific population hinders the required validation of the combination methodology.[11]

2. Validation is possible

The availability of data gives the cardiologist some additional options to improve the risk assessment. In fact, current tools can be adjusted for that population or as an alternative the hospital can develop a tool directly from that data containing a specific set of risk factors. The performance of risk assessment tools can be evaluated for their selection or to define the respective importance (weights) in a potential combination scheme.

The methodology explored in this dissertation also takes advantage of the availability of data. The performance of individual models assessed from the data is used to guide the individual models selection procedure as well as to define the respective weights to incorporate the combination scheme. An additional optimization procedure can be carried out to adjust the performance of the global model to that specific population.

It is important to refer that the availability of data may also permit the derivation of a specific model for that population eliminating the need of the proposed combination methodology. However, this hypothetical new model would be one more risk assessment tool derived based on a specific population. The proposed strategy is more flexible as it merges knowledge provided by different current tools that are known and accepted by the physicians and simultaneously assures the elimination of some of the identified weaknesses of those tools.

[11] Here, the only possible validation procedure relies on the comparison between the risk assessment provided by the model with the technical opinion of the cardiologist.

Personalization Based on Grouping of Patients

The introduction of some personalization issues in risk assessment can originate important benefits, namely the CVD event risk assessment performance improvement exclusively based on the proper selection of available risk assessment tools. This particular aspect can be identified as another important contribution of this thesis.

The reduction of the false positive and false negative cases[12] is mandatory in order to increase the utilization of the risk assessment tools within the daily clinical practice context.

1.4 Clinical Support

The main goals of this thesis were validated by the leader of the Cardiology Department of Leiria-Pombal Hospital Centre, Portugal. This close collaboration was essential as it assures that the main contributions of this research work are clinically relevant. The clinical partner confirmed that:

- According to the international guidelines all patients must have their risk evaluated. Risk scores must be used in clinical practice since they have an incontestable clinical significance;
- The classification in two categories (low risk/high risk) is correct. In fact, the aim of cardiologist in clinical practice is to discriminate between high risk patients and low risk patients. From a clinical perspective, the identification of intermediate risk patients is not very significant;
- Physicians are aware that risk score models have different accuracies depending on the specific test situations;
- Having several tools, the physician may have some difficulty to define the weights of the different individual models to combine (weighted combination). It is easier for physicians to define the weights/importance of individual variables;
- False negative errors are more important than false positive errors;
- There are some important limitations of the current risk models:
  - They are not able/have difficulty to adapt to specific populations;

[12] False positive: patients with a positive diagnosis who were incorrectly diagnosed; False negative: patients with a negative diagnosis who were incorrectly diagnosed.

- They consider a limited number of variables. So, models are not able to incorporate variables that can be as important as those that were used to develop the model;
- They are not able/have difficulty to cope with missing information;
- They do not allow the direct incorporation of clinical knowledge that physicians collect in the daily clinical practice.

Two other clinical collaborations, Castle Hill Hospital (Hull, UK) and Santa Cruz Hospital (Lisbon, Portugal), also played an important role in validating the main targets of this thesis as well as helping to obtain the required real patient data for validation purposes.

1.5 Structure

This dissertation intends to provide a complete description of the proposed methodologies as well as of all the performed validation procedures and corresponding conclusions. It can be systematized as follows:

Chapter 2 provides relevant background to the most important issues in this thesis. It contains information about the more suitable risk assessment tools regarding the patient's conditions (CAD, HF) under analysis. A comparison between classifiers is done to clarify the selection of Bayesian classifiers in order to implement the common representation of individual risk assessment tools. In this context, probabilistic classifiers, namely the naïve Bayes classifier, are detailed. Several combination methodologies of individual models are also explored. They can be organized in two different categories: *i*) models' output combination; *ii*) models' fusion. Given their specific properties, genetic algorithms (GA) were selected among the optimization algorithms to improve the global classifier's performance. Moreover some techniques that deal with missing risk factors are identified. Dimensionality reduction techniques as well as clustering algorithms are explored as they were applied to the implementation of the personalization based on grouping of patients' strategy. Validation is a critical phase of the proposed approach. Several validation issues are explored in this chapter with some focus on bootstrapping validation and statistical significance tests.

Chapter 3 presents the developed methodologies in this work. The common representation based on naïve Bayes classifier is detailed. The selection of that

specific classifier is clarified as well as the description of the parameter's learning procedure required for its implementation. The weighted average combination scheme is introduced. The optimization procedure based on genetic algorithms operation is also described. The combination methodology was validated in two different situations: *i*) simulation – theoretical individual models; *ii*) tools applied in the clinical practice. The approach that can be applied to incorporate clinical knowledge is also explored. Finally, the personalization based on grouping of patients is also depicted as well as the respective validation strategy.

Chapter 4 contains the results that were originated through the different validation tests that were performed throughout this work. The two mentioned validation scenarios[13] of combination methodology were addressed and the corresponding validation results are shown separately. The results of the incorporation of clinical knowledge are also presented. Additionally, the validation results from the personalization based on grouping of patients' methodology are depicted.

Chapter 5 is the final chapter of this dissertation. The obtained results are discussed and the main conclusions are reached. Future potential developments/improvements of the proposed approach are pointed out. In addition, the list of scientific publications produced during this thesis is detailed.

[13] The results obtained in the two distinct validation scenarios were based on different datasets.

2. Background

2.1 Introduction

This chapter provides relevant background to the most important issues in this thesis. In order to provide a global perspective of the issues addressed its structure is presented in Figure 2.1:

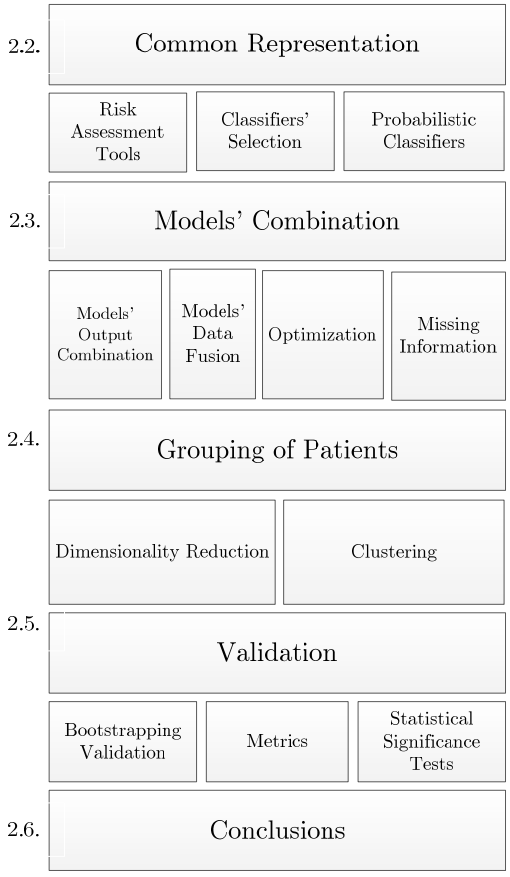


Figure 2.1 - Structure of chapter 2.

As previously mentioned, current risk assessment tools are diversely represented (charts, equations, etc.) which hinders their combination. This diversity of representations is not suitable to deal with missing risk factors nor is it able to incorporate additional clinical knowledge. Therefore, the first step of the methodology presented in Figure 1.2 is to create a common representation of individual risk assessment tools that minimizes the identified weaknesses. This topic is addressed in Section 2.2. The first part of the section is dedicated to the description of current available CVD risk assessment tools, namely the identification of the most relevant tools according to the considered patient's conditions (CAD, HF). Some techniques to derive these risk assessment tools are described and their flaws are also identified. The second part of the section is dedicated to the selection of the machine learning classifier to implement that common representation. Several classifiers are described and compared through the identification of their advantages and disadvantages. Afterwards, Bayesian classifiers are detailed. Some Bayesian network concepts (structure and parameterization, inference mechanism, learning methods) are also addressed. The clarification of these issues is important in order to introduce the specific features of the Bayesian classifiers, namely the naïve Bayes classifier.

The second step of the methodology (Figure 1.2) is the combination of individual models. This combination has several advantages, as it: *i*) avoids the discarding of the available information originated by the previously developed tools; *ii*) allows the consideration of a higher number of risk factors; *iii*) avoids the choice of a "standard model". Section 2.3 explores this issue based on two different approaches: *i*) models' output combination; *ii*) models' parameter/data fusion. For both categories, several current methods to implement the combination of models are explored along with some relevant research work in this topic. Optimization algorithms, particularly the genetic algorithms, are also depicted. In fact, the adjustment of parameters in the global model is an important feature of the proposed methodology. Finally, the global model's ability to deal with missing risk factors is one of the main requirements of the methodology developed in this work. For that reason, some techniques that deal with missing risk factors in machine learning context are also described.

An additional methodology to improve the risk assessment performance when compared to the one obtained with the current risk assessment tools is also proposed in this work. This methodology is based on the personalization of the risk prediction through patients' grouping. In this context dimensionality reduction techniques

together with unsupervised learning algorithms[14], namely clustering algorithms are central to its implementation. Section 2.4 presents an overview of dimensionality reduction methods as well as some relevant topics of clustering algorithms.

Validation assumes a critical importance to assure that the developed methodology has some potential to be applied in the clinical practice. Section 2.5 depicts some topics that are relevant for the definition of the validation procedure. In this section, different types of validation are identified, closely viewed is the bootstrapping validation due to its importance for this research. The second part of the section details some metrics that are frequently applied to evaluate the performance of machine learning classifiers. In the last part statistical hypothesis tests are explored as they are an important tool in the implemented validation procedure.

Section 2.6 gathers the techniques referred that were applied in this work.

2.2 Common Representation

2.2.1 Cardiovascular Risk Assessment Tools

As mentioned, prognosis[15] of cardiovascular disease relates to the probability or risk of an individual to develop a particular outcome over a specific period of time. This assessment is based on both clinical/non-clinical information that is available at the time of the prediction *(Moons, 2009)*

CVD risk assessment tools consider a specific period of time (long term (years)/short term (months), and they differ on: input risk factors, particular type of cardiovascular disease (coronary artery disease, heart failure, etc.), events/end point (fatal/non-fatal), prevention type (primary/secondary) and patient's specific condition (diabetics, CAD, HF, etc.).

These tools may also be grouped according to the prevention type. In fact, there are tools specific to patients with established cardiovascular disease (secondary

[14] Find hidden structures in unlabeled data.

[15] Prognosis differs from diagnosis as the latter is related with the determination of the possibility of a disease from current symptoms, signs and tests *(Visweswaran, 2007)*.

prevention) and others intended to assess CVD risk on patients who have not yet established the disease (primary prevention) *(Hobbs, 2004)*.

Primary Prevention

There are several long term tools suitable for primary prevention. These tools differ on the considered risk factors, specific disease, event type and period of time considered for the prediction.

Table 2.1 presents the main features of some of the most relevant risk score tools for primary prevention:

Model	Patients Enrolled	Disease/Event	Term (years)	Patient's condition	Risk Factors
Framingham *(D'Agostino, 2008)**	8491	CVD	10		Age, Sex, TC, HDL, SBP, SMK, DB, BPT
Joint British Societies *(JBS, 2005)*	n.a.	CHD/Death	10		Age, Sex, TC, HDL, SBP, SMK, DB, BPT
PROCAM *(Assmann, 2002)*	5389	CHD	10		Age, Sex, TC, HDL, SBP, SMK, DB, BPT, FH, TRG
QRISK *(Cox, 2007)*	1.28 million	CVD (Heart attack / Stroke)	10		Age, Sex, TC, HDL, SBP, SMK, DB, BPT, FH, BMI, PE, KD, ETH, RHU
SCORE *(Conroy, 2003)*	205178	CVD/Death	10		Age, Sex, TC, HDL, SBP, SMK, BPT
Sheffield *(Wallis, 2000)*	1000	CHD	10		Age, Sex, TC, HDL, SBP, LVH, SMK, DB, BPT
UKPDS *(Stevens, 2001)*	4540	CHD	1-20	Diabetics	Age, Sex, TC, HDL, SBP, SMK, BPT, HE, NY
ASSIGN *(Woodward, 2007)*	13297	CVD	10		Age, Sex, TC, HDL, SBP, DB, SIM, CPD, FH

*11th biennial examination cycle of original cohort (1968 to 1971)

CVD – Cardiovascular disease, **CAD/CHD** – Coronary heart disease, **HF** – Heart failure, **TC** – Total cholesterol, **BPT** – Blood pressure treatment, **FH**- Family history, **TRG** – Triglycerides, **HE**- Haemoglobin, **NY** – n^{o} years diagnosis, **CPD** – Cigarettes per day, **SIM** - Social deprivation Index, **LVH** – Left ventricular hypertrophy, **SMK** – Smoking, **DB** – Diabetes**, PE** – Previous event, **KD** – Kidney disease, **ETH** – Ethnicity, **RHU** – Rheumatoid, **SBP** – Systolic blood pressure, **HDL** – High density lipoprotein.

Table 2.1 - Primary prevention: risk assessment tools.

Secondary Prevention

Secondary prevention risk assessment tools are specific to patients with established cardiovascular disease.

Due to their social and economic impact, two specific patient conditions are taken into account: *i*) heart failure (HF); *ii*) coronary artery disease (CAD).

1. Heart Failure

Heart failure can be defined as the failure of the heart to pump blood with normal efficiency. When this occurs, the heart is unable to provide adequate blood flow to other organs such as the brain, liver and kidneys. Heart failure may be due to failure of the right or left or both ventricles *(WHO, 2007).*

According to Swedberg *(Swedberg, 2005)*, heart failure is a syndrome in which the patients should have the following features: symptoms of heart failure, typically breathlessness or fatigue, either at rest or during exertion, or ankle swelling and objective evidence of cardiac dysfunction at rest.

The term acute heart failure (AHF) is often adopted to designate a decompensation of chronic heart failure (CHF) characterized by signs of pulmonary congestion, including pulmonary edema.

Several risk assessment tools specific to HF disease can be identified, since HF is a chronic disease with an adverse prognosis given that it presents a high one-year mortality rate 35% to 40% *(Lee, 2003)*. These systems differ on predicted events (death, develop HF, etc.), input risk factors, patient conditions (ambulatory patients, hospitalized patients, cardiac transplant candidates, etc.), period of time (months/years).

Table 2.2 depicts the main features of some of the most well known risk score systems, as well as their input risk factors:

Model	Patients Enrolled	Event	Term (months)	Patients' Condition	Risk Factors
EFFECT *(Lee, 2003)*	2624	Death	1/12	HF Hospitalized HF Ambulatory	Age, DM, CNR, CVA, CRR, CLD, SBP, DBP, RR, BUN, SD
Bouvy *(Bouvy, 2003)*	152	Death	18	HF Ambulatory HF Hospitalized	Age, Sex, DB, RKD, WT, SBP, DBP, AO, BBK
Adlam *(Adlam, 2005)*	532	Death	60	CAD/Primary Care	Age, Sex, DB,STK, ECGA,BNP
ABC *(Butler, 2008)*	2935	Develop HF	60	CAD/Primary Care	Age, SK, CAD, SBP, LVH, RHR, GL, CR, AL
Senni *(Senni, 2006)*	807	Death	12	HF Ambulatory	Age, NYHA, VHD, DB, RKD, CNR, COPD, SBP, DBP, LVEF, AF, HE, BBK, ACE
Rich *(Rich, 2006)*	282	Death	60	HF Ambulatory	Age, CAD, DM, PAD, SBP, DBP; SD; BUN
SHFM *(Levy, 2006)*	1125	Death	12, 60	HF Ambulatory	Age, Sex, HT, NYHA, WT, SBP, DBP, LVEF, IMI, SD, CHL, HE, LY, UA, BK, ACE, ARB, ST, KSD
Kannel *(Kannel, 1999)*	486	Develop HF	48	CAD	Age, Sex, HT, VHD, CAD, DB, WT, SBP, DBP, LVH, RHR
ADHERE *(Fonarow, 2005)*	32229	Death	6	HF Hospitalized HF Ambulatory	SBP, BUN, CR
HFSS *(Aaronson, 1997)*	268	Death Urgent transplant	1/12	HF Ambulatory	CAD, SBP, DBP, LVEF, IVCD, POC, RHR, SD
Charm *(Pocock, 2005)*	7599	Death HF hosp.	12	HF Ambulatory	Age, Sex, HT, NYHA, SK, CAD, DB, DR, HL6, RCR, WT, SBP, DBP, PE, AE, LVEF, IMI, BBB, CGL, RHR, INS
Brophy *(Brophy, 2004)*	4277	Death	24	HF Ambulatory	Age, NYHA, DE, LA, RCR, S3, SBP, DBP, LVEF, CDR, CR, NT
MUSIC *(Vasquez, 2009)*	992	Death	44	HF Ambulatory	ASVD; LAS; NSVT/PVC; AF; BNP; TR;HYP; EGFR

DM – Dementia; **CVA**- Cerebrovascular accident; **CLD** - Chronic lung disease; **CRR** – Cirrhosis, **SD** - Sodium, **DB** – Diabetes, **RKD** - Renal/kidney dysfunction, **AO -** Ankle edema, **BK-** BBlocker, **CNR** – Cancer**, WT**- weight**, SBP** - Systolic blood pressure; **DBP** - Diastolic blood pressure; **STK** – Stroke, **ECGA** – ECG abnormalities, **SK-** Smoking, **RR** - Respiratory rate; **RHR** - Resting heart rate; **GL** - Glucose**, CR** – Creatinine**, AL** – Albumin**, HE** – Hemoglobin, **HT** – Height, **IMI** - Ischemic/MI, **CHL** – Cholesterol, **LY -** Lymphocytes**, UA** – Uric acid, **ST** – Statin, **KSD** - K-sparing diuretic, **BUN** - Blood urea nitrogen; **BNP** – B-type natriuretic peptide ; **CAD** - Coronary artery disease; **LVH** - Left ventricular hypertrophy; **COPD**-Chronic obstructive pulmonary; **LVEF** – Left ventricular ejection fraction; **AF** – Atrial fibrillation; **VHD** – Valvular heart disease; **PAD** - Peripheral artery disease; **ARB** – Angiotensin receptor blocker; **ACE** – Angiotensin converting enzyme inhibitor; **RHR** – Resting heart rate ; **RCR** - Rales/crackles **IVCD** – Intraventricular conduction delay; **POC** – Peak oxygen consumption; **DR** – Dyspnea at rest; **HL6** – Hospitalization last 6 months ; **PE** – Pulmonary edema; **AE** – Ankle edema; **CGL** – Cardiomegaly, **INS** – Insulin, **S3** - S3 gallop**, NT -** Nitrates, **MR** – Mitral regurgitation; **BBB** - Bundle branch block; **DE** – Dyspnea at exercise; **LA** - Limitation of activity; **CDR** - Cardiothoracic ratio; **ASVD** - Atherosclerotic vascular event; **LAS** - Left atrial size, **NSVT/PVC** – Non sustained ventricular tachycardia/Premature ventricular contraction; **EGFR** - Estimated glomerular filtration Rate, **TR** – Troponin, **HYP** – Hyponatremia

Table 2.2 – Secondary prevention: risk assessment tools for heart failure.

2. Coronary Artery Disease

Coronary artery disease begins when hard cholesterol substances (plaques) are deposited within coronary arteries that ensure the supply of blood rich in oxygen and nutrients to the heart. Coronary arteries begin at the base of the aorta and spread across the surface of the heart, branching out to all areas of the heart muscle *(WHO, 2007)*. Plaques deposited in coronary arteries can originate a clot that may reduce or even stop the flow of blood to the heart muscle. If coronary arteries become too narrow, the blood supply to the heart muscle is reduced causing chest pain (angina pectoris). Heart attack or myocardial infarction occurs when a plaque ruptures originating a blood clot that obstructs the artery and stops the blood flow to part of the heart muscle. That part of the heart muscle dies *(WHO, 2007)*. As stated, CAD is the single most important cause of death in Europe.

Some tools are developed for secondary prevention and specific to patients with coronary artery disease. Due to the severity of this disease, these risk assessment tools predict the risk of an event in a short period of time (months). These systems differ on predicted events (death, myocardial infarction, urgent revascularization, etc.), input risk factors, period of time (days, months). Table 2.3 presents the main features of some of the most well known risk score systems specific for CAD patients, as well as their input risk factors:

Model	Patients Enrolled	Event	Term (months)	Patient's condition	Risk Factors
GRACE *(Tang, 2007)*	1143	Death/MI	6	CAD	Age, SBP, CAA HR, CR, STD, ECE, KIL
PURSUIT *(Boersma, 2000)*	337	Death	1	CAD	Age, Sex, SBP, CCS, HR, STD, ERL, HF
TIMI NSTEMI *(Antman, 2000)*	3171	Death/MI/ UR	14 days	CAD	Age, STD, ECE, KCAD, ASP, ANG, RF
TIMI STEMI *(Morrow, 2008)*	14114	Death	1	CAD	Age, SBP, HR, CHF, DB, HYP, ANG, WT, ASTE, LBBB, RX4

MI – Myocardial infarction, **UR** – Urgent revascularization; **SBP** – Systolic blood pressure**, CR** - Creatinine, **HR** – Heart rate, **CAA** – Cardiac arrest at admission, **KIL** –Killip class: II-IV, **STD** - ST segment depression, **ECE** - Elevated cardiac enzymes, **KCAD**- Known coronary artery disease, **ERL** – Enrolment(MI/UA), **HF** – Heart failure, **CCS** – Angina classification, **ASP** - Use of aspirin in the previous 7 days, **ANG** - 2 or more angina events in past 24 h, **RF** - 3 or more cardiac risk factors, **DB** – Diabetes, **HYP** – Hypertension, **WT** – Weight, **ASTE** -Anterior ST segment elevation, **LBBB** - Left bundle branch block, **RX4** – Time to treatment >4 hours

Table 2.3 - Secondary prevention: risk assessment tools for coronary artery disease.

2.2.2 Risk Assessment Tools' Derivation

Statistical Techniques

Usually, CVD risk assessment tools are derived based on statistical techniques that require the close monitoring of population samples over a long period of time (Cui, 2009). Survival analysis is the area of statistics that is used to analyze the survival time of the patients in a clinical study *(Rossi, 2010)*. The main goal of survival analysis is to identify the relationship between survival time, time that an event takes to occur, and one or more predictors (risk factors). Namely, survival analysis allows the building of a model for the survival probability based on a collection of variables (predictors) that are believed to influence the survival time of an individual *(Rossi, 2010) (Ata, 2007)*.

There are two important functions in survival analysis that must be depicted: survival function and hazard function (hazard rate).

Survival function represents the probability of survival up to time t, where T is a random variable that represents survival time. Survival function $S(t)$ is always a decreasing function that starts with value 1 *(Rossi, 2010)*:

$$S(t) = P(T > t) \tag{2.1}$$

Hazard rate $h(t)$ assesses the instantaneous risk of death (event) at time t of a patient, given that the patient survived up to that time:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < (t + \Delta t \mid T \ge t)}{\Delta t} \tag{2.2}$$

These two functions are closely related, by equation (2.3) where $f(t)$ denotes the death density function:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.3}$$

One of the most common methods in survival analysis is the Cox proportional hazard model, which is given by:

$$h(t \mid \mathbf{x}) = h_0(t) \times e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p} \tag{2.4}$$

where $\mathbf{x} = [x_1 \ ... \ x_p]^T$ is the set of observed values for the p predictors, $\boldsymbol{\beta} = [\beta_1 \ ... \ \beta_p]^T$ are the unknown parameters of the model, called proportional hazard regression coefficients, $h_0(t)$ is the baseline hazard function, that is the value of $h(t)$ obtained when all the predictors assume the value 0.

Based on the regression coefficient calculations, it is possible to define the probability of an event occurrence, the absolute risk, as:

$$P_{event} = 1 - S_t^{\exp(\boldsymbol{\beta}^T(\mathbf{x}-\bar{\mathbf{x}}))} \tag{2.5}$$

The parameter S_t is the value of the survival function at the end of the considered period of time for analysis and $\bar{\mathbf{x}}$ is the vector of the mean values of the p predictors.

Machine Learning Algorithms

The statistical methods are the most frequently adopted by the health care research community. However machine learning algorithms have also been applied to derive CVD risk assessment tools.

Voss derived a model based on PROCAM study using neural networks (5115 men aged 35-65 years at recruitment). According to the author *(Voss, 2002)*, a multilayer perceptron neural network improved the risk prediction when compared to standard logistic regression. Ning *(Ning, 2006)* also applied neural networks to develop a method to stratify the cardiovascular risk among hypertension patients. This study considered 348 subjects, 269 hypertensive and 79 normotensive patients. The obtained results confirmed the accuracy of the model and demonstrated its ability to risk assessment for patients with hypertension. Valanavis *(Valavanis, 2010)* adopted neural networks to perform a multifactorial analysis of obesity as a CVD risk factor. A model to predict obesity was implemented based on 2341 patients. The obtained results confirmed the potential of neural networks to build the desired predictive model.

Support vector machines (SVM) were also used to assess the risk of cardiovascular disease. Alty *(Alty, 2007)* developed an approach based on SVM to predict CVD considering the evaluation of the arterial stiffness[16]. Some features of the

[16] Arterial stiffness (loss of elasticity of the arteries) causes the arteriosclerosis.

Digital Volume Pulse (DVP)[17] waveform were extracted to estimate the arterial stiffness. Then, SVM were applied to stratify the CVD risk, achieving high accuracy within a population of 461 patients. Several applications of support vector machines to predict risk (risk stratification) were developed in the clinical area, e.g. *(Kasamatsu, 2008)*, *(Balasubramanian, 2009)*.

These approaches based on neural networks and support vector machines are accurate but are *black box* models, which inhibits the clinical interpretability of the risk assessment model.

Decision tree is another classifier used to predict cardiovascular risk. Ordonez *(Ordonez, 2006)* compared the performance of association rules[18] and decision trees for cardiovascular disease prediction, concluding that decision trees are less effective than constrained association rules. Ture *(Ture, 2005)* compared decision rules with neural networks so as to predict the development of hypertension among a population of 694 subjects. Decision trees performed worse than neural networks. Although, rather than neural networks the decision trees assure the interpretability of the prediction model.

Some risk prediction systems are created based on Bayesian networks. Nicholson *(Nicholson, 2008)* created two models, one based on the Busselton study (8000 participants) and the other on the PROCAM study. In this study, Bayesian networks had a similar performance to the logistic regression models for assessing the CAD risk. Other research works centered in the exploitation of the potential of Bayesian networks were developed, e.g. *(Verduijn, 2007)*, *(Atoui, 2006)*. Bayesian networks show the causal relationships between variables and assure the interpretability of the prediction model.

Pitt *(Pitt, 2009)* performed a very comprehensive comparison among some of the referred machine learning classifiers. The aim of the study was to identify new risk factors associated with anesthesia procedures that may influence the cardiovascular risk namely, to assess the cardiovascular risk associated with anesthesia delivery as well as to identify the most appropriate anesthetic agent.

---

[17] The digital volume pulse (DVP) is recorded by measuring the transmission of infra-red light absorbed through the finger. DVP varies with red blood cell density, its amplitude depends on temperature and perfusion of the hand and its contour is related with characteristics of the heart and large arteries.

[18] Association rule learning is a method for discovering relations between variables in large databases. Association rules exhaustively look for hidden patterns, which can be applied for discovering predictive rules involving subsets of the medical data set attributes.

Machine learning classifiers are important tools in the implementation of risk assessment models. Section 2.2.3 depicts some of the features of the most important machine learning classifiers along with their advantages/disadvantages.

Weaknesses of Current CVD Risk Tools

As mentioned, risk assessment tools are important to help physicians in their daily practice. However, these tools present some flaws that are important to emphasize:

- Individually they only consider a very limited number of risk factors. This aspect forces physicians to make a prognosis based on a restricted set of available information;
- Current risk assessment models are not prepared to deal with missing risk factors. Although, missing information is a very frequent problem both in daily clinical practice and in health records *(Khanna, 2005)*;
- They do not allow the incorporation of clinical knowledge. Empirical clinical knowledge is a key factor that should be ideally included in risk models to allow a better prognosis;
- Dynamics of risk evolution is another issue that is not addressed by current risk tools. It is important to evaluate the risk variation due to changes in risk factors. For instance, changes in smoking habits must be properly considered in risk assessment;
- Lack of personalization also hinders the operation of current risk assessment tools. In fact, they were derived from a set of known patient's and they are applied to new patients' cases. Visweswaran *(Visweswaran, 2007)* designates these models as *population-wide methods*. He states that *patient-specific models* perform better than *population-wide methods* when they are applied to a particular patient case.

This dissertation aims to circumvent some of these problems. Thus, the common representation of individual risk assessment tools (Figure 1.2) should be represented based on a proper supervised machine learning classifier such that: *i*) facilitates the combination of individual models; *ii*) should be able to deal with missing risk factors; *iii*) allows the incorporation of additional knowledge.

2.2.3 Supervised Machine Learning Classifier's Selection

Machine learning is the process of using observations to build a model that can predict a new observation *(Ulrich, 2008)*, e.g. determine the CVD risk of a new patient. Machine learning can either be supervised or unsupervised, depending on whether labeled training data or unlabeled training data is supplied. Kotsiantis *(Kotsiantis, 2007)* systematized supervised machine learning classifiers in five categories: logic based algorithms, perceptron based techniques, probabilistic reasoning, instance based learning and support vector machines.

Logic Based Algorithms

Decision trees and learning set of rules are included in logic based algorithms, as they classify instances based on decision nodes or decision rules. Instances' classification is done by sorting them out based on feature values creating a set/hierarchy of tests. It is important to stress that decision trees can be translated into a set of rules by creating a separate rule for each path from the root to a leaf in the tree. However, rules can also be directly induced from a training data using a variety of rule-based algorithms *(Murthy, 1988) (Furnkranz, 1999)*.

1. Decision Trees

A decision tree is a hierarchical model (Figure 2.2), composed of nodes, branches and leaves. A decision node d_i tests an attribute, each branch b_{ij} corresponds to an attribute value and a leaf node L_i assigns a classification *(Dwyer, 2007)*.

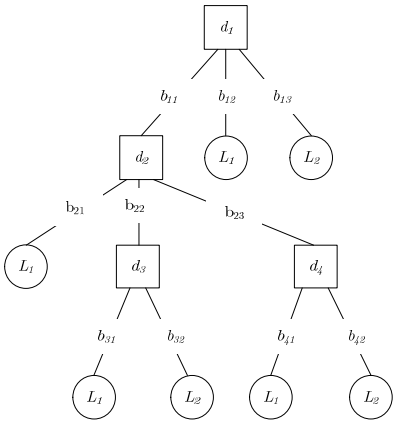


Figure 2.2 - Example of a decision tree structure.

A decision tree defines a function $f:\Upsilon \to C$, where Υ is a set of instances (input data) and C is a finite set of classes (output values). ψ is the set of attributes and it is defined by the particular learning problem. Each instance[19] $\mathbf{x} \in \Upsilon$ consists of a set of attribute-value pairs.

Each decision node d_i contains a test on a specific attribute $x \in \psi$. A specific decision node originates q_i disjoint branches (e.g. d_1 originates $q_1 = 3$ branches), such that the decision node connects to its q_i descendant (nodes or leaves) through q_i different branches. When a decision node is defined, a split is performed in the tested attribute, which means that smaller size subsets have been created from the original set of instances Υ. A leaf node L_i has no branches but has a specific label $c_i \in C = \{c_1,...,c_m\}$ which matches the respective class output.

There are some specific decision tree algorithms to define the *goodness measure*[20] of the candidate splits as well as the stopping criteria to discontinue further splitting. One of the most well-known algorithms is C4.5 *(Quinlan, 1992)*, that uses the gain criterion as *goodness measure*. The gain criterion is an information based measure that is defined based on the different proportions of the decision test outcomes, such that:

$$I(D) = -\sum_{i=1}^{m} \mathrm{p}_i \times log_2(\mathrm{p}_i) \tag{2.6}$$

where m denotes the number of classes, p_i denotes the fraction of instances that belong to the output class i. Given $D = \{(\mathbf{x}_1, c_1),...,(\mathbf{x}_N, c_N)\}$ the gain criterion that results from the split v is determined as follows:

$$gain_I(D,v) = I(D) - \sum_{i=1}^{q} \frac{|D_i|}{|D|} \times I(D_i) \tag{2.7}$$

q is the number of outcomes of split v and $|D_i|/|D|$ is the fraction of instances in the node that belong to the subset D_i. The best candidate split is given by *(Quinlan, 1992) (Mingers, 1989)*:

---

[19] If $\mathbf{x}_i$ is known to belong to a particular class pair $(\mathbf{x}_i, c)$ it is designated for labeled instance, if the output class is unknown it is designated by unlabeled instance.

[20] It is used in Decision Trees to refer the quality of the splitting process. The goodness measure ranks the candidate splits and it is a key aspect in the tree-growing process.

$$v^{*} = \underset{v \in V}{argmax} \quad gain(D, v) \tag{2.8}$$

In order to prevent over-fitting[21] prune operation is required. There are two types of pruning: pre-pruning when the growing phase is stopped prematurely and post-pruning which means that portions of the tree are removed after it has been grown. Algorithm C4.5 implements post-pruning namely a technique designated by *pessimistic pruning*. An existing subtree is replaced by a leaf whenever the leaf has a lower predicted error rate *(Furnkranz, 1999) (Quinlan, 1992)*.

It is important to stress that there are other algorithms that implement their own criteria to induce the decision tree.

Decision trees could be a valid option to implement the common representation of individual risk assessment tools due to their accuracy and interpretability. However, their lack of ability to deal with missing risk factors obstructs this possibility.

2. Learning Set of Rules

Learning set of rules is one of the most expressive and human readable supervised machine learning classifiers *(Mitchell, 1997)*, and it is based on *if-then* rules:

$$if \quad A_1 \wedge ... \wedge A_n \quad then \quad Q$$

where $\{A_1,...,A_n\}$ are designated for clause antecedents, Q for clause consequent and the operator $\wedge$ denotes the logic conjunction of the clause antecedents.

The goal is to implement the smallest set of rules that covers the whole training data set. Usually a large number of learned rules means that the learning algorithm is able to *remember* the training dataset rather than discover the relationships that contribute to its structure, which leads to over-fitting *(Kotsiantis, 2007)*.

A set of rules can be obtained from a decision tree or directly induced from a training data set. *Separate-and-conquer* algorithms derive rules directly from the training data set. The algorithm searches for a rule that covers part of the training instances, removes those instances and repeats the process (conquer) on the remaining examples. The process iterates until no examples remain, which assures

[21] The algorithm loses its capability of generalization, i.e. it makes poor predictions on unseen cases.

that each instance is covered by at least one rule. There are several *separate-and-conquer* algorithms that differ on the method to learn single rules, on rule evaluation and on procedures to avoid over-fitting.

According to Furnkranz *(Furnkranz, 1999)*, this type of algorithm can be detailed as presented in the Figure 2.3:

```
SeparateAndConquer (instances)
theory = ∅
while Positive (instances) ≠ ∅
    rule=FindBestRule (instances)
    covered=Cover(rule, instances)
    If RulesStoppingCriterion (theory, rule, instances)
            exit
    instances = instances – covered
    theory = theory ∪ rule
return (theory)
```

Figure 2.3 - Generic separate and conquer algorithm (Furnkranz, 1999).

The function *FindBestRule* is used for learning a rule by maximizing an evaluation rule criterion (heuristic function). Covered data instances are separated from the training data set, the learned rule is stored and another rule is learned from the remaining examples. The process iterates until the stop condition is reached or there are no more remaining instances.

The order of rules is a key aspect to the operation of these classifiers in multi-class classification problems. In these situations each instance can be covered by several rules. Different rule orders can originate different predictions of the output class for the same instance.

Similarly to decision trees, learning set of rules provides a high interpretability of the model although it has difficulty to cope with missing risk factors.

Perceptron based techniques

Perceptron based classifiers (Artificial Neural Networks - ANN) are an important class of classifiers that are applied to several real world classification problems *(Zhang, 2000).*

Figure 2.4 presents the perceptron neuron model which is the basic processing unit of a multilayer perceptron neural network.

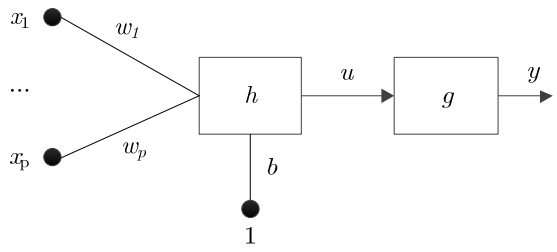


Figure 2.4 - Perceptron neuron model (Gonzaléz, 2008).

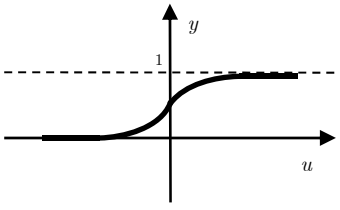
where $\mathbf{x} = [x_1,...,x_p]^T$ are the inputs, y is the output, b is designated the bias, the weight vector is represented as $\mathbf{w} = [w_1,...,w_p]^T$, h is the combination function and g is the activation function. Given that $b \in R$; $\mathbf{w} \in R^p$ the output u of the combination function h is given by:

$$u = h(\mathbf{x}, b, \mathbf{w}) = b + \sum_{i=1}^{p} w_i x_i \tag{2.9}$$

while output y is derived using the activation function g:

$$y = g(\mathbf{x}, b, \mathbf{w}) = g(b + \sum_{i=1}^{p} w_i x_i) \tag{2.10}$$

There are several activation functions, however the threshold function, the linear function and the sigmoid function can be identified as the most commonly used in perceptron models *(Gonzaléz, 2008) (Demuth, 2002).*



$$y = g(u) = \frac{1}{1 + e^{-u}}$$

Figure 2.5 - Sigmoid function (Demuth, 2002).

Neurons can be combined to form a neural network. The architecture of a neural network is defined by the neurons' number and by the way that they are connected *(Gonzaléz, 2008).*

Multilayer perceptron is a neural network with a feed-forward architecture, which is presented in the Figure 2.6:

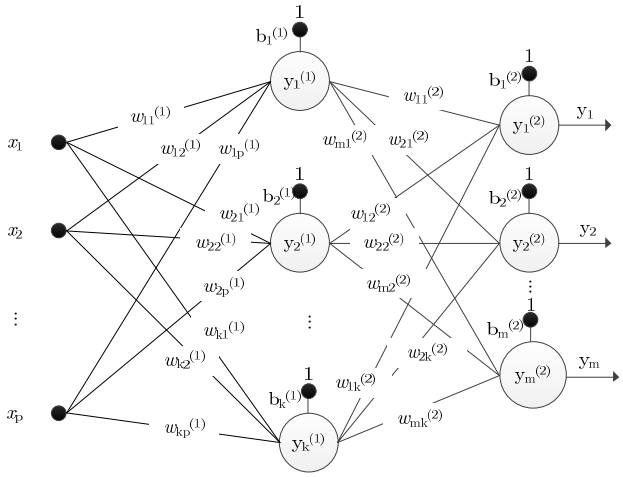


Figure 2.6 -Multilayer perceptron (Gonzalez, 2008).

A feed forward architecture can be represented as an acyclic graph. Neurons are grouped in $(z+1)$ layers, z hidden layers $L^{(1)}...L^{(z)}$ plus the output layer $L^{(z+1)}$.

In this architecture neurons that belong to one layer feed the next layer, so communication is done layer by layer, i.e. starting from the input layer all the way through the hidden layers to the output layer. The outputs of the output layer are the results of the neural network computation *(Sıma, 2003)*.

Considering a two layer perceptron, Figure 2.6, the output computation is performed as follows:

$$\begin{aligned} u_j^{(1)} &= b_j^{(1)} + \sum_{i=1}^{p} w_{ji}^{(1)} x_i \\ y_j^{(1)} &= g^{(1)}(u_j^{(1)}) \end{aligned} \tag{2.11}$$

Where $g^{(1)}$ is the activation function of the hidden layer. Outputs of the hidden layer are the inputs of the output layer; k is the number of neurons of the hidden layer:

$$\begin{aligned} u_j^{(2)} &= b_j^{(2)} + \sum_{i=1}^{k} w_{ji}^{(2)} y_i^{(1)} \\ y_j^{(2)} &= g^{(2)}(u_j^{(2)}) \end{aligned} \tag{2.12}$$

A multilayer perceptron with one hidden layer containing the sigmoid activation function and an output layer comprising of the linear activation function is a universal approximator. In fact, it can approximate any function from one finite dimensional space R^p to another R^m. The approximation accuracy depends on the number of neurons that constitute the hidden layer of the neural network *(Hornik, 1991)*.

The behaviour of neural networks depends on three aspects: network architecture; activation functions and weights of each input connection *(Kotsiantis, 2007)*. Weights are determined through a learning process based on a training dataset.

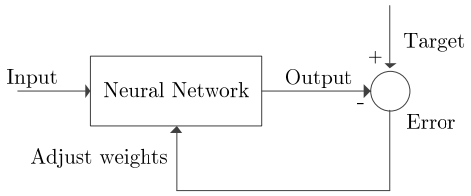


Figure 2.7 - Training process.

There are several training algorithms for multilayer perceptron to perform the learning process, e.g. back propagation, momentum, variable learning rate, genetic algorithms, Newton's method and other back propagation variations *(Neocleous, 2002). (Yam, 2002) (Vivarelli, 2001) (Parekh, 2000) (Whitley, 1995).*

It is also important to refer, that there are different neural networks' architectures like radial basis function (RBF) networks that are also universal approximators. However, this issue will not be covered in this thesis.

As referred, artificial neural networks are powerful universal approximators that in this case have the potential to perform an accurate reproduction of the individual risk assessment tools' behavior. Nevertheless, these classifiers do not assure the interpretability of the model nor do they have the ability to deal with missing risk factors.

Instance Based Learning

Instance based learning classifiers are lazy-learning algorithms, which means that they delay the induction of the model until the classification is required[22]. Instance

[22] This is the opposite of eager learning where the model is induced from the training data set before the classification is necessary *(Duda, 2000) (Friedman, 1996).*

based learning algorithms consist of storing a set of training examples (training data set) and when a new instance is encountered, a set of similar related instances is retrieved from memory and used to classify the query instance (target function). Thus less computation effort is required during the training phase while more is used during the classification process.

One of the most well-known instance based learning algorithm is the k-nearest neighbour (kNN). The kNN algorithm classifies an instance through the majority vote of its neighbours, with the instance being assigned to the class most voted amongst its k nearest neighbours. If k is set to one, then the instance is simply assigned to the class of its nearest neighbour *(Duda, 2000)*.

In order to identify the nearest neighbours a distance function is applied (Table 2.4) *(Wilson, 2000)*. The distance function sorts the training instances in relation to the query instance and k determines how many instances are selected and used as neighbours *(Wang, 2006)*. The distance/similarity function and the choice of k are the key aspects defining the kNN performance.

Euclidean	$d(\mathbf{u},\mathbf{v}) = \sqrt{\sum_{i=1}^{p}(u_i - v_i)^2}$
Minkowsky	$d(\mathbf{u},\mathbf{v}) = \sqrt[r]{\sum_{i=1}^{p}\left\|u_i - v_i\right\|^r}$
Manhattan	$d(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{p}\left\|u_i - v_i\right\|$
Chebychev	$d(\mathbf{u},\mathbf{v}) = \max_{i=1}^{p}\left\|u_i - v_i\right\|$
Camberra	$d(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{p}\frac{\left\|u_i - v_i\right\|}{\left\|u_i + v_i\right\|}$
Kendall's Rank Correlation	$d(\mathbf{u},\mathbf{v}) = 1 - \frac{2}{p(p-1)!}\sum_{i=1}^{p}\sum_{j=1}^{i-1} sig(u_i - u_j)sig(v_i - v_j)$ $\quad sig(x) = \begin{cases} -1; & if\ x < 0 \\ 0; & if\ x = 0 \\ 1; & if\ x > 0 \end{cases}$

$\mathbf{u} = [u_1 ... u_p]^T$; $\mathbf{v} = [v_1 ... v_p]^T$; p number of attributes

Table 2.4 - Distance between instances kNN algorithm (Wilson, 2000).

Usually classification is done by voting among the selected neighbours:

$$f : R^p \rightarrow C$$
$$c_{\mathbf{x}_q} \leftarrow \underset{c \in C}{argmax} \sum_{i=1}^{k} \delta(c, c_{\mathbf{x}_i}) \quad \delta(a,b) = \begin{cases} 1; \ if \ a = b \\ 0; \ if \ a \neq b \end{cases} \tag{2.13}$$

$\mathbf{x}_1,...,\mathbf{x}_k$ denote the k training-instances that are nearest to the query instance $\mathbf{x}_q$ and $c_{\mathbf{x}_i} \in C$ the output class of instance $\mathbf{x}_i$. The vote of each neighbour can also be weighted considering its distance to the query instance. As expected, weights increase proportionally to the reduction of the distance between selected neighbours and the query instance as defined in (2.14).

$$\begin{cases} w_i = \dfrac{d(\mathbf{x}_k, \mathbf{x}_q) - d(\mathbf{x}_i, \mathbf{x}_q)}{d(\mathbf{x}_k, \mathbf{x}_q) - d(\mathbf{x}_1, \mathbf{x}_q)} & d(\mathbf{x}_k, \mathbf{x}_q) \neq d(\mathbf{x}_1, \mathbf{x}_q) \\ w_i = 1; & d(\mathbf{x}_k, \mathbf{x}_q) = d(\mathbf{x}_1, \mathbf{x}_q) \end{cases} \tag{2.14}$$

where $\mathbf{x}_1,...,\mathbf{x}_k$ are the k nearest neighbours of $\mathbf{x}_q$ arranged in increasing order of $d(\mathbf{x}_i, \mathbf{x}_q)$, therefore $\mathbf{x}_1$ is the closest neighbour of $\mathbf{x}_q$.

The expression of $c_{\mathbf{x}_q}$ must be updated to incorporate the weight factor:

$$f : R^p \rightarrow C$$
$$c_{\mathbf{x}_q} \leftarrow \underset{c \in C}{argmax} \sum_{i=1}^{k} w_i \delta(c, c_{\mathbf{x}_i}) \quad \delta(a,b) = \begin{cases} 1; \ if \ a = b \\ 0; \ if \ a \neq b \end{cases} \tag{2.15}$$

There are several algorithms to define the weights of the nearest neighbors *(Wettschereck, 1997)*. There is also a significant number of algorithms to implement other instance based learning classifiers, even though several of those algorithms are variants of the *kNN* algorithm *(Brighton, 2002) (Wilson, 2000) (Aha, 1991)*.

Support Vector Machines

Support Vector Machine (SVM) is a class of algorithms that performs the classification task based on the determination of a hyperplane within a multidimensional space. This hyperplane can separate instances with different class labels and it can be applied to both linear and nonlinear separable data *(Kotsiantis, 2007)*.

The equation of a hyperplane that separates positive from negative labelled linear separable instances represented as $(\mathbf{x}_i, c_i)$; $\mathbf{x}_i \in R^p$, $c_i \in \{-1,1\}$ can be defined as:

$$\boldsymbol{w}.\mathbf{x} + b = 0 \tag{2.16}$$

where $\boldsymbol{w} = [w_1,...,w_p]; \mathbf{x} = [x_1,...,x_p]^T$, $\boldsymbol{w}.\mathbf{x}$ is the dot product, $\boldsymbol{w}$ is a vector normal to the hyperplane, $b/\|\boldsymbol{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\boldsymbol{w}\|$ is the Euclidean norm of $\boldsymbol{w}$ (Figure 2.8) *(Burges, 1998)*. The decision rule is given by:

$$f_{\boldsymbol{w},b}(\mathbf{x}) = sign(\boldsymbol{w}.\mathbf{x} + b) \tag{2.17}$$

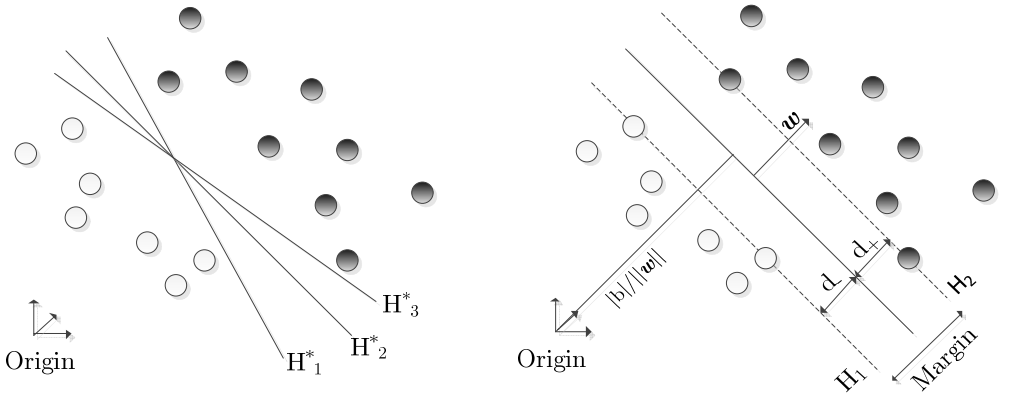


Figure 2.8 – Hyperplane - separable cases (Burges, 1998).

It is possible to define several hyperplanes to separate the same training instances (Figure 2.8 a). The best hyperplane (optimal hyperplane) must be chosen such that a small shift in the data should not result in prediction changes. Actually, if the distance between hyperplane and training instances is minor, e.g. H_1^*, H_3^* in (Figure 2.8 a), test examples that are very close to training examples can be incorrectly classified *(Burges, 1998)(Lukas, 2003)*. Therefore, the distance between the hyperplane and the nearest training instances must be maximized in order to optimize the generalization capability of the classification algorithm. The concepts of

separability and *margin* must be depicted to support the optimal hyperplane definition.

The hyperplane defined by $\boldsymbol{w}$ and b is called a separating hyperplane if:

$$\begin{aligned} &\mathbf{x}_i.\boldsymbol{w}+b \geq +1 \;\; for \;\; c_i = +1 \\ &\mathbf{x}_i.\boldsymbol{w}+b \leq -1 \;\; for \;\; c_i = -1 \end{aligned} \tag{2.18}$$

The margin $\xi_k(\boldsymbol{w},b)$ of a training instance $\mathbf{x}_k$ is defined as the distance between the hyperplane and $\mathbf{x}_k$:

$$\xi_k(\boldsymbol{w},b) = c_k(\boldsymbol{w}.\mathbf{x}_k + b) \tag{2.19}$$

The margin of a set of r instances $\Upsilon_r = \{\mathbf{x}_1,...,\mathbf{x}_r\}$ is defined as:

$$\xi_{\Upsilon_r}(\boldsymbol{w},b) = \min_{\mathbf{x}_k \in \Upsilon_r} \xi_k(\boldsymbol{w},b) \tag{2.20}$$

Figure 2.8 b) presents the graphical description of this margin concept, which is given by $(d_+ + d_-)$ where d_+ is the distance to the closest positive instance and d_- the distance to the closest negative instance. For the linear separable case the optimal hyperplane is the one that assures the largest margin, and is defined by:

$$(\boldsymbol{w}^*, b^*) = \underset{\boldsymbol{w},b}{argmax} \;\; \xi_\Upsilon(\boldsymbol{w},b) \tag{2.21}$$

The optimal hyperplane must be calculated through the following optimization problem:

$$\begin{aligned} &min \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &restricted \;\; to: \\ &\mathbf{x}_i.\boldsymbol{w}+b \geq +1 \;\; for \;\; c_i = +1 \\ &\mathbf{x}_i.\boldsymbol{w}+b \leq -1 \;\; for \;\; c_i = -1 \end{aligned} \tag{2.22}$$

The calculation of the optimal hyperplane for the linear nonseparable case, must consider that some misclassifications must be tolerated in the overlapping region *(Lukas, 2003)*, since the linear separation of the training instances is not possible. However, each violation of the optimization problem constraints originates a misclassification penalty *(Lukas, 2003) (Cortes, 1995)*. Thus, the equation (2.19) is modified to:

$$c_k[\boldsymbol{w}.\mathbf{x}_k + b] \geq 1 - \zeta_k \tag{2.23}$$

where ζ_k is given by:

$$\zeta_k = max\left\{0;1-c_k[\boldsymbol{w}.\mathbf{x}_k+b]\right\} \tag{2.24}$$

and it measures the instances misclassification; $\zeta_k > 1$ means that $\mathbf{x}_k$ is misclassified; $0 < \zeta_k < 1$ means that $\mathbf{x}_k$ is correctly classified but inside the margin and $\zeta_k = 0$ indicates $\mathbf{x}_k$ is correctly classified outside the margin.

Based on the consideration of ζ_k, the optimization problem becomes:

$$\begin{aligned} &\min_{\boldsymbol{w},b,\zeta} \frac{1}{2}\left\|\boldsymbol{w}\right\|^2 + \gamma\sum_{i=1}^{N}\zeta_i \\ &restricted\ to: \\ &c_k[\boldsymbol{w}.\mathbf{x}_k+b] \geq 1-\zeta_k \qquad k=1,...,N \\ &\zeta_k > 0 \end{aligned} \tag{2.25}$$

where γ is a positive constant, N is the number of instances.

A more detailed description of this classifier is beyond the scope of this thesis, however a very comprehensive survey on support vector machine classifiers can be found in *(Campbell, 2002), (Burges, 1998).*

In spite of this classifier's high accuracy, its lack of interpretability along with its difficulty to deal with missing risk factor makes it inappropriate for the common representation of individual risk assessment tools.

Probabilistic Reasoning

Rather than a deterministic classification, probabilistic/statistical learning algorithms provide a probability that an instance belongs to each class.

Bayesian network is a probabilistic model that combines a graphical representation (structure) with quantitative information (parameters/conditional probabilities) to represent a joint probability distribution over a set of random variables[23]. Due to their flexibility and causality representation, these networks have been frequently used within the medical field, for diagnosis, patient monitoring and therapy planning *(Visweswaran, 2007) (Roberts, 2006).*

---

[23] Given a probability space (ψ, P) a random variable X is a function on ψ. A random variable assigns a unique value to each element in the sample space, creating a set of values called the space of X. A random variable is discrete if its space is countable. The joint probability distribution of two (or more) random variables X_1, X_2 defined on the sample space ψ is given by $P(X_1, X_2)$ *(Neapolitan, 2004).*

Bayesian classifiers are probabilistic classifiers that implement particular structures of Bayesian networks, as their goal is to assign a class label to instances described by a given set of attributes. Classification relies on Bayes rule to predict the class of C with the highest probability given the value of an attribute X (Friedman, 1997).

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \tag{2.26}$$

Usually X is an observation (e.g., clinical exam) and C a hypothesis (e.g., have a disease). The term $P(C \mid X)$ denotes a posterior probability, i.e., the probability of the hypothesis after having seen the observation X (probability to have a disease given the results of a clinical exam). $P(C)$ is the prior belief, the probability of the hypothesis before seeing any observation (prevalence of the disease). $P(X \mid C)$ is a likelihood, the probability of the observation if the hypothesis is true (sensitivity of the clinical exam).

An important classifier, naïve Bayes (Figure 2.9), assumes a particular configuration of a Bayesian network, which is composed of a directed acyclic graph (DAG) with only one parent (unobserved node) and several children (observed nodes).

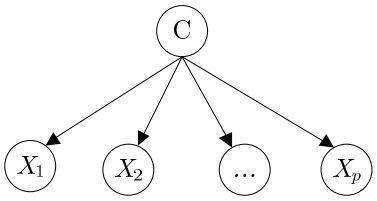


Figure 2.9 - Naïve Bayes structure.

Naïve Bayes classifier has a key importance in this dissertation, since it was selected to implement the common representation of the individual risk assessment tools. Therefore, probabilistic reasoning classifiers namely naïve Bayes classifier will be detailed in Section 2.2.4.

Machine Learning Classifiers Comparison

All the mentioned classifiers have strengths and weaknesses. Thus, it is important to systematize the main requirements that the candidate model must verify to implement the common representation of risk assessment tools: *i*)

interpretability of the model. This feature is mandatory not only to enable the combination of current risk assessment tools but also to facilitate the incorporation of additional clinical knowledge; *ii*) ability to deal with missing risk factors. As stated the incapacity to cope with missing risk factors is one of the identified weaknesses of the current risk assessment tools that must be overcome; *iii*) competitive performance with other machine learning classifiers. Moreover, Kotsiantis *(Kotsiantis, 2007)* made a comparison between these five categories of machine learning classifiers, as presented in Table 2.5.

	Classifier	Advantages	Disadvantages
a	Decision Trees	Interpretability	Difficulty to deal with missing information There is no common accepted algorithm to build DT Requires pruning
	Learning Set of Rules	Interpretability	Difficulty to deal with missing information
b	Neural Networks	Accuracy	No interpretability Incapability to deal with missing information Easily leads to over-fitting
c	k-nearest neighbor (kNN)	Simple	Incapability to deal with missing information Large effort for classification High sensitivity to different similarity functions
d	Support Vector Machines	Suitable when the number of features is larger than the number of training instances	Difficulty to deal with missing information;
e	Naïve Bayes	Simplicity; Interpretability Able to deal with missing information	Attributes' independence assumption

a)Logic Based Algorithms; b) Perceptron Based Techniques; c) Instance Based Learning; d) Support Vector Machines; e) Probabilistic.

Table 2.5 - Classifiers comparison (Kotsiantis, 2007).

Considering the specific requirements of the model to implement the common representation of individual risk assessment tools together with the data of Table 2.5, the naïve Bayes classifier was selected. The naïve Bayes has a competitive performance with remaining classifiers, is simple and can deal with missing risk factors. Besides these features, naïve Bayes assures the interpretability of the model which is critical to allow the implementation of the proposed combination methodology. Finally, the structure of naïve Bayes simplifies the incorporation of empirical clinical knowledge.

The identified characteristics make this classifier particularly suitable for the approach developed in this work. Probabilistic classifiers, namely the naïve Bayes classifier, are comprehensively detailed in the following section.

2.2.4 Probabilistic Classifiers

Due to their importance for the proposed approach, Bayesian classifiers are detailed. Firstly, some important Bayesian network concepts (structure and parameterization, inference mechanism, learning methods) are explored. Then, Bayesian classifiers, in particular naïve Bayes classifier, are described.

Bayesian Networks

The structure of a Bayesian network is defined through a Directed Acyclic Graph[24] (DAG) as presented in Figure 2.10:

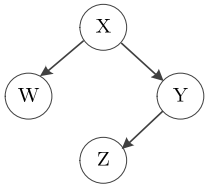


Figure 2.10 - Directed acyclic graph (Roberts, 2006).

Assuming the notation defined by Neapolitan *(Neapolitan, 2004),* a directed graph is a pair (V,E) where V is a finite nonempty set whose elements are the nodes, and E is a set of ordered pairs of distinct elements of V designated by edges (arcs). If $\{X,Y\} \in V$ and $(X,Y) \in E$, then there is an edge from X to Y. A directed acyclic graph does not contain any cycle.

Given a DAG $G = (V,E)$ and nodes X,Y:

- X is called a parent of Y (child) if there is an edge from X to Y.
- If there is a path from X to Y, Y is called a descendant of X and X is designated as an ancestor of Y.

[24] A graph can be described as a set of nodes that are connected by a set of edges. The edges can be either directed (arrows) or undirected, depending on whether they point from one node to another or simply indicate a link between nodes.

- If there is no path from X to Y, Y is called a non-descendant of X.
- If there is an edge from X to Y or from Y to X, these nodes are adjacent.
- A path from X_1 to X_k is the set of edges that connects the k nodes $\{X_1, X_2, ..., X_k\} \in V;\ k \geq 2$.

The parameterization $\theta_{\mathcal{G}}$ is a set of local probabilistic models that quantitatively encode the dependence of each variable on its parents. There is a local probability distribution defined on each node X_i that considers each state of its parents. This conditional probability distribution $P(X_i \mid Pa_i)$[25] depends on the type of the variables involved (continuous, discrete) and on the specific relationship between variables *(Visweswaran, 2007).* If the random variables are discrete, $P(X_i \mid Pa_i)$ is presented as a table that contains a cell for each joint instantiation of $P(X_i = x_i \mid Pa_i = pa_i)$[26] where x_i represents the values that X_i may assume and pa_i the different possible states of the respective X_i's parents. Each row (column) in the table, called a conditional probability table (CPT), represents a single conditional probability distribution $P(X_i \mid Pa_i = pa_i)$.

In spite of its straightforward implementation, the CPT can have some drawbacks: *i*) the number of parameters grows exponentially with the number of parents Pa_i as well as with their possible states. The increase of the number of states of Pa_i can also lead to a poor estimate of the CPT parameters; *ii*) the tabular representation ignores the potential interaction between the parents of X_i. This aspect may increase the number of parameters needed to specify the conditional probability distribution of the variables, e.g. parents' independence:

$$P(X \mid Y, Z = z) = P(X \mid Z = z) \tag{2.27}$$

for all values of X, Y when $Z = z$. A more comprehensive approach to this topic can be found in *(Visweswaran, 2007) (Neapolitan, 2004).*

Therefore, a Bayesian network may be described as proposed by Cooper *(Cooper, 1999)* in Figure 2.11:

- The nodes represent variables of interest which may be discrete or continuous.

---

[25] Pa_i: set of nodes that are parents of X_i.

[26] In a Bayesian network random variables are denoted by capital letters while lower-case letters are used for the values that these variables can assume.

- The set of directed links represent the conditional dependencies among the variables.
- The strength of an influence is represented by conditional probabilities attached to each cluster of parent-child nodes in the network (Table 2.6).

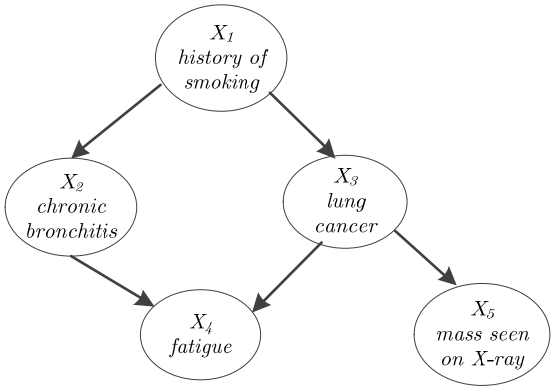


Figure 2.11 - Example of a Bayesian network (Cooper, 1999).

$P(X_1 = no) = 0.8$	$P(X_1 = yes) = 0.2$
$P(X_2 = absent \mid X_1 = no) = 0.95$	$P(X_2 = present \mid X_1 = no) = 0.05$
$P(X_2 = absent \mid X_1 = yes) = 0.75$	$P(X_2 = present \mid X_1 = yes) = 0.25$
$P(X_3 = absent \mid X_3 = absent) = 0.99995$	$P(X_3 = present \mid X_3 = absent) = 0.00005$
$P(X_3 = absent \mid X_3 = present) = 0.997$	$P(X_3 = present \mid X_3 = present) = 0.003$
$P(X_4 = absent \mid X_2 = absent, X_3 = absent) = 0.95$	$P(X_4 = present \mid X_2 = absent, X_3 = absent) = 0.05$
$P(X_4 = absent \mid X_2 = absent, X_3 = present) = 0.5$	$P(X_4 = present \mid X_2 = absent, X_3 = present) = 0.5$
$P(X_4 = absent \mid X_2 = present, X_3 = absent) = 0.9$	$P(X_4 = present \mid X_2 = present, X_3 = absent) = 0.1$
$P(X_4 = absent \mid X_2 = present, X_3 = present) = 0.25$	$P(X_4 = present \mid X_2 = present, X_3 = present) = 0.75$
$P(X_5 = absent \mid X_3 = absent) = 0.98$	$P(X_5 = present \mid X_3 = absent) = 0.02$
$P(X_5 = absent \mid X_3 = present) = 0.4$	$P(X_5 = present \mid X_3 = present) = 0.6$

Table 2.6 - Conditional probabilities table (Cooper, 1999).

A Bayesian network structure $\mathcal{G}$ encodes the set of independencies among the variables in the domain that are defined based on the application of local and global Markov conditions. The awareness of these Markov conditions is a key aspect in understanding the Bayesian network's operation.

Local Markov condition states that: *a node is conditionally independent of its non-descendants given the state of its parents.* If a node does not have parents the node is simply independent of its non-descendants *(Cooper, 1999) (Niculescu, 2005).* This condition translates a high dimensional multivariate joint probability distribution into a product of potentially low dimensional probability distributions *(Visweswaran, 2007).* This aspect can be formalized in the following way:

- Let the variables $\{X_1, X_2, ..., X_n\}$ be the nodes of $\mathcal{G}$, that are arranged such that if $i < j$ then X_i is a non-descendant of X_j in $\mathcal{G}$:

  Applying the chain rule of probability the joint probability of $\{X_1, X_2, ..., X_n\}$ is:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, ..., X_{i-1}) \tag{2.28}$$

  The local Markov condition states that for all $X_i \in \{X_1, X_2, ..., X_n\}$:

$$P(X_i \mid X_1, ..., X_n) = P(X_i \mid Pa_i) \tag{2.29}$$

  where $Pa_i \subseteq \{X_1, ..., X_{i-1}\}$. According to the ordered arrangement of variables all of the parents of X_i are in the set of $\{X_1, ..., X_{i-1}\}$ and none of the descendants of X_i are in this set. Then, the chain rule for Bayesian networks is obtained based on equation (2.30):

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid Pa_i) \tag{2.30}$$

Considering the Bayesian network presented in Figure 2.11, the joint probability of $\{X_1, X_2, X_3, X_4, X_5\}$ can be calculated (2.31) through equation (2.30).

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2, X_3)P(X_5 \mid X_3) \tag{2.31}$$

The global Markov condition states that: *a node is conditionally independent of all other nodes in the network, given its parents, its children and the children's parents (Neapolitan, 2004).* This set of nodes is known as the Markov blanket. The application of this global condition enables the identification of all conditional independencies.

The concept of *d-separation* captures all the conditional independence relationships that occur in a Bayesian network. In fact, *d-separation* extends the Markov conditions to the identification of independencies among disjoint sets of nodes. Given three disjoint subsets of nodes X_s, Y_s, Z_s in structure $\mathcal{G}$, X_s is independent of Y_s given Z_s if nodes in Z_s block all the existing paths between nodes of X_s and nodes in Y_s. In this way, it is possible to identify the independencies among groups of nodes.

1. Bayesian Inference

Through equation (2.30), it is possible to calculate any joint probability, e.g. the probability $P(X_1 = yes, X_2 = present, X_3 = present, X_4 = present, X_5 = present)$ (Figure 2.11; Table 2.6) is calculated as follows:

$$\begin{aligned} &P(X_1 = yes)\ P(X_2 = present \mid X_1 = yes)\ P(X_3 = present \mid X_1 = yes) \\ &P(X_4 = present \mid X_2 = present, X_3 = present)\ P(X_5 = present \mid X_3 = present) \\ &= 0.2 \times 0.25 \times 0.003 \times 0.75 \times 0.6 \\ &= 0.0000675 \end{aligned} \tag{2.32}$$

However, the capability of inference of a Bayesian network is often more useful than the calculation of a joint probability. The inference mechanism in a Bayesian network intends to derive the posterior probability of one or more variables given the values observed for other variables. Based on a simple example, Cooper *(Cooper, 1999)* describes this mechanism very clearly: Let S_1 and S_2 be sets of variables with assigned values, the inference mechanism should provide the value of $P(S_1 \mid S_2)$. For instance, in Figure 2.11 S_1 is $X_1 = yes$ and S_2 assumes $X_4 = present$:

$$\begin{aligned} &P(S_1 \mid S_2) = P(X_1 = yes \mid X_4 = present) \\ &P(S_1 \mid S_2) = \frac{P(S_1 \cap S_2)}{P(S_2)} \end{aligned} \tag{2.33}$$

This equation implies the determination of $P(S_1 \cap S_2)$, which requires the sum over all the combinations of value assignments to the variables that do not belong to the considered sets $(S_1; S_2)$.

$$P(S_1) = P(X_1 = yes)\ ;\ P(S_2) = P(X_4 = present)$$

$$\begin{aligned}
&P(S_1 \cap S_2) = \\
&P(X_1 = yes)\ P(X_2 = present \mid X_1 = yes)\ P(X_3 = present \mid X_1 = yes) \\
&P(X_4 = present \mid X_2 = present, X_3 = present)\ P(X_5 = present \mid X_3 = present) \\
&+ \\
&P(X_1{=}yes)\ P(X_2{=}absent|X_1{=}yes)\ P(X_3{=}absent|X_1{=}yes) \\
&P(X_4 = present \mid X_2 = absent, X_3 = absent)\ P(X_5 = absent \mid X_3 = absent) \\
&+ \\
&...
\end{aligned} \tag{2.34}$$

A similar reasoning can be implemented to the calculation of $P(S_2)$:

$$P(S_2) = P(X_4 = present)$$

$$\begin{aligned}
&P(S_2) = \\
&P(X_1 = no)\ P(X_2 = present \mid X_1 = no)\ P(X_3 = present \mid X_1 = no) \\
&P(X_4 = present \mid X_2 = present, X_3 = present)\ P(X_5 = present \mid X_3 = present) \\
&+ \\
&P(X_1{=}yes)\ P(X_2{=}absent|X_1{=}yes)\ P(X_3{=}absent|X_1{=}yes) \\
&P(X_4 = present \mid X_2 = absent, X_3 = absent)\ P(X_5 = absent \mid X_3 = absent) \\
&+ \\
&...
\end{aligned} \tag{2.35}$$

This direct inference mechanism has a major problem, as its time complexity is exponential with the number of variables (nodes) present in the network. For large networks equation (2.33) may be unfeasible *(Cooper, 1999)*.

Some algorithms were developed to circumvent this problem. Although, there is no algorithm that can be efficiently[27] applied to all Bayesian networks *(Cooper, 1999)*. According to Niculescu *(Niculescu, 2005)*, inference in a Bayesian network is a *NP-hard* problem that originated the development of several algorithms such as: variable elimination, message passing on junction trees, Markov chain Monte Carlo, etc.

The exhaustive knowledge of these inference algorithms is beyond the scope of this thesis, however there are several scientific publications available that detail the Bayesian networks inference process, e.g. *(Neapolitan, 2004) (Boutilier, 1996) (Jordan, 1998)*.

[27] This term refers to computational efficiency.

2. Learning Bayesian Networks

This is other important aspect to understand the construction process of a Bayesian network. A Bayesian network can be derived from prior experience (experts/ data available in literature) or through the learning of the model structure and distributions from real data. Alternatively, the two approaches can be combined.

Several algorithms were developed to build Bayesian networks directly from data *(Roberts, 2006)*, *(Niculescu, 2005), (Neapolitan, 2004), (Heckerman, 1999)*. They can be focused on two main issues: *i*) structure definition; *ii*) parameter estimation.

Structure Definition

A Bayesian network structure $\mathcal{G}$ encodes the relationships that are established among the several nodes X_i that belong to the network. Therefore, the structure definition allows the identification of the dependencies and independencies among the domain variables. If different Bayesian networks can represent the same distributions, such structures are said to be Markov equivalents[28] *(Cooper, 1999)*.

There are two major approaches for learning the structure of Bayesian networks:

- Constraint-based methods that employ statistical independence tests among the domain variables, to determine the presence or absence of arcs in the network. The final Bayesian network is the one that best represents the relationships between variables. The accuracy of these tests can be seriously affected by the eventual lack of data;
- Search and score methods that apply a metric (Table 2.7) to evaluate the goodness of fit of the statistical model represented by a specific structure.

  These search and score methods evaluate how well the corresponding statistical model fits the observed data. Heckerman *(Heckerman, 1999)* states that given a scoring function, a training data set and a space of possible network structures, the goal of search and score method is to find a network structure that maximize that score. Finding the highest-scoring network can be a NP-hard problem which allows the utilization of heuristic techniques.

[28] Two Bayesian network structures are Markov equivalent if and only if they contain the same set of variables and they represent the same conditional independence relationships on those variables, as given by the Markov condition *(Cooper, 1999)*.

Likelihood Score	$$score_L(\mathcal{G};D)=\sum_{i=1}^{n}\sum_{j=1}^{qi}\sum_{k=1}^{ri}N_{ijk}\log\frac{N_{ijk}}{N_{ij}}$$ n is the number of nodes, q_i is the number of states that parents Pa_i of node X_i may assume, r_i is the number of values that X_i can take, N_{ijk} is the number of cases in the training data such that $X_i = k \wedge Pa_i = j$ and $N_{ij} = \sum_k N_{ijk}$; $D = \{(\mathbf{x}_1, c_1),...,(\mathbf{x}_N, c_N)\}$ is the dataset.
Description length score	$$score_{DL}(\mathcal{G};D)=\frac{Dim[\mathcal{G}]}{2}\log N - score_L(\mathcal{G};D)$$ $Dim(\mathcal{G})$ is the number of parameters and N is the cardinality of data D.
Bayesian score	$$score_B(\mathcal{G};D)=\log P(D\mid\mathcal{G})+\log P(\mathcal{G})$$ The prior $P(\mathcal{G})$ assigns prior probabilities for different graph structures, $P(D \mid \mathcal{G})$ measures the goodness of fit of the given structure to the data.

Table 2.7 - Scoring functions (Visweswaran, 2007).

Bayesian networks structure learning methods are explored in several works *(Cooper, 1999), (Heckerman, 1999), (Neapolitan, 2004), (Visweswaran, 2007).*

Structure learning is not applied in this work as the common representation of individual risk assessment tools is implemented through a naïve Bayes classifier which has a specific structure (Figure 2.9).

Parameter Estimation

The goal of parameter estimation is to find the proper values for each parameter (conditional probabilities) in the Bayesian network. According to Visweswaran *(Visweswaran, 2007)* this learning procedure can be formulated as:

- The parameter learning can be described as a hypothesis space which defines the set of all possible values being considered and a scoring function that scores different hypothesis in the space relative to the given data. This learning process is achieved assuming that Bayesian network structure is known, all the variables are discrete and the data has no missing values.

Two main learning methods can be identified: *i*) maximum likelihood estimation; *ii*) Bayesian parameter estimation.

The maximum likelihood estimation is a frequentist approach that tries to estimate the "best set" of parameters θ *(Niculescu, 2005)*, i.e., it measures how well the different possible parameters' values predict the data.

Considering a Bayesian network with n discrete nodes, the parameterization over the entire network is $\theta = \{\theta_1,...,\theta_n\}$, where θ_i represents the conditional probability table $P(X_i \mid Pa_i)$ associated with the node X_i. Each θ_i is decomposed as $\theta_i = \{\theta_{i1},...,\theta_{ij},...,\theta_{iq_i}\}$ where q_i is the number of possible instantiations of Pa_i. Each θ_{ij} represents the parameters defining the single conditional distribution $P(X_i \mid Pa_i = j)$. This distribution is the multinomial distribution such that:

$$P(X_i \mid Pa_i = j) = multinomial(\theta_{ij}) \tag{2.36}$$

θ_{ij} can be further decomposed as $\theta_{ij} = \{\theta_{ij1,...,}\theta_{ijk,...,}\theta_{ijn_i}\}$ [29] where $\theta_{ijk} = P(X_i = k \mid Pa_i = j)$ and n_i is the number of possible instantiations of X_i. Usually, it is assumed that θ_i and θ_{ij} are mutually independent which is designated respectively as global and local parameter independency. The maximum likelihood estimator $\hat{\theta}_{ijk}$ parameters for θ_{ij} are determined through:

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}} \tag{2.37}$$

where N_{ijk} is the number of cases in the training dataset such that $X_i = k$ and $Pa_i = j$. The total number of instances in the training data set must verify the equation (2.38).

$$N_{ij} = \sum_k N_{ijk} \tag{2.38}$$

[29] θ_{ijk} is the set of parameters of the Conditional probability table $P(X_i \mid Pa_i)$, where rows are associated with the different values of $X_i = k$ and columns comprise the values of $Pa_i = j$.

McLachlan *(Mclachlan, 2008)* presents the Expectation maximization algorithm that allows the maximum likelihood estimation even when the data is not fully observable (missing/hidden data).

Contrarily to the frequentist approach, Bayesian parameter estimation does not intend to find a single set of parameters $\hat{\theta}$ that explain the data, but it provides a distribution over the possible parameters' value that quantifies the uncertainty of each of the values *(Visweswaran, 2007)*. Several choices of parameters are possible but some choices have a higher prior probability to occur *(Niculescu, 2005)*. Thus, the Bayesian approach combines prior knowledge on the parameters θ with the posterior distribution of a new data D given θ.

The Bayes rule is applied to compute:

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \tag{2.39}$$

where $P(D \mid \theta)$ is the likelihood function, $P(\theta)$ is the prior distribution of θ which represents the prior knowledge/belief about the different values of the parameters. and $P(D)$ is the marginal likelihood that represents a priori likelihood of observing the obtained data given the prior belief.

$$P(D) = \int_{\theta} P(D \mid \theta)P(\theta)d\theta \tag{2.40}$$

A more detailed description of the Bayesian parameter estimation can be found in *(Niculescu, 2005)*.

Bayesian Classifiers

As already mentioned, Bayesian classifiers are probabilistic classifiers that implement particular structures of Bayesian networks, as their goal is to assign a class label to instances described by a given set of attributes. Classification relies on Bayes rule (2.26) to predict the class of C with the highest probability given a set of attributes $\{X_1,...,X_p\}$ *(Friedman, 1997)*.

It is important to clarify the following notation:

- $\mathbf{X} = [X_1,...,X_p]$ is a vector of random variables that contains the p observed attributes;

- $\mathbf{x} = [x_1,...,x_p]$ a particular instance that contains the observed values of the different p attributes;
- $\mathbf{X} = \mathbf{x}$ the same as $X_1 = x_1 \wedge X_2 = x_2 \wedge ... \wedge X_p = x_p$;
- C is a random variable that denotes the class of an instance;
- c is a particular class label.

1. Naïve Bayes Classifier

Naïve Bayes classifier implements a particular structure of a Bayesian network, which is represented in Figure 2.12.

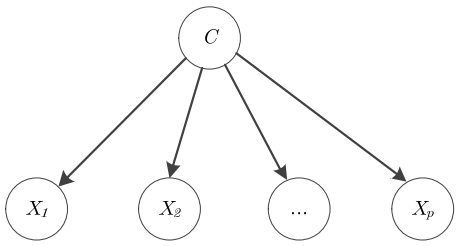


Figure 2.12 - Naïve Bayes structure.

This classifier is composed of only one parent (unobserved node: C) and several children (observed nodes: $X_1,...,X_p$). Despite its simple configuration, naïve Bayes is computationally efficient, can deal with incomplete information and presents a predictive performance competitive with other classifiers *(Friedman, 1997) (Tsymbal, 2003) (Twardy, 2005).*

Its structure imposes a strong independence condition: all the attributes X_i are conditionally independent[30] given the value of the class C. Frequently the conditional independence between attributes is unrealistic, however even if this condition is not verified, naïve Bayes often presents a performance comparable to other classifiers *(Friedman, 1997) (Tsymbal, 2003).*

Consider a graph structure $\mathcal{G}$ comprising the class variable that is the root node[31] and a set of discrete random variables (attributes) $\{X_1,...,X_p\}$ where each attribute has the class variable as its unique parent, namely $Pa_i = \{C\}$ for all

[30] Probabilistic Independence: A is independent of B given C whenever $P(A \mid B,C) = P(A \mid C)$ for all possible values of A,B,C whenever $P(C) > 0$.

[31] A node that has no parents.

$1 \le i \le p$. Based on (2.30) the joint probability of the variables that belong to $\mathcal{G}$ is given by:

$$P(X_1,...,X_p,C) = P(C)\prod_{i=1}^{p} P(X_i \mid C) \tag{2.41}$$

Based on the definition of conditional probability[32], $P(C \mid \mathbf{X})$ is estimated according to the following equation:

$$P(C \mid \mathbf{X}) = P(C \mid X_1,...,X_p) = \alpha P(C)\prod_{i=1}^{p} P(X_i \mid C) \tag{2.42}$$

where α is a normalization constant *(Friedman, 1997)*. Then the final classification c is achieved based on the following equation:

$$c = \underset{c_j}{argmax}(\alpha P(c_j)\prod_{i=1}^{p} P(x_i \mid c_j)) \tag{2.43}$$

c_j is a mutually exclusive class of C, x_i is the value of attribute X_i that belongs to the query instance $\mathbf{x}_q = [x_1,...,x_p]$ and $\alpha = 1/P(\mathbf{X} = \mathbf{x}_q)$.

The structure of naïve Bayes classifier is completely defined, therefore the learning process is related only with model parameters. The model has to learn from the training data set, the conditional probability $P(X_i \mid C)$ of each attribute X_i given the class C as well as the prior probability $P(C)$ of the class C. Here, the discretization of numeric attributes has great impact in the construction of the conditional probability table and therefore in the performance of the classifier.

Conditional Probabilities Calculation

After the discretization procedure is concluded[33], conditional probability tables must be calculated. This calculation for a training data set with N instances is given by:

---

[32] $\mathrm{P(X,Y)} = \mathrm{P(X \cap Y)};\ \mathrm{P(X \mid Y)} = \dfrac{\mathrm{P(X,Y)}}{\mathrm{P(Y)}}$

[33] Discretization methods are explored in the next section.

$$P(X_i = x_i \mid C = c) = \frac{\sum^{N}(X_i = x_i \wedge C = c)}{\sum^{N}(C = c)} \tag{2.44}$$

the variable $c \in C$ that has several categories $C = \{c_1,...,c_n\}$ (mutually exclusive), the variable x_i denotes a particular value of the attribute X_i, N is the total number of instances.

2. Discretization Methods

An attribute X_i can be either qualitative (categorical) or quantitative (numeric). The different nature of the attributes has a critical importance in the probabilities estimation *(Yang, 2009)*.

Qualitative data usually assumes a small number of possible values, which allows a reliable estimation of probabilities: *i*) $P(C = c)$ that can be estimated from the frequency of instances with $C = c$; *ii*) $P(X_i = x_i \mid C = c)$ that can be estimated from the frequency of instances $X_i = x_i \wedge C = c$ considering the total of instances $C = c$.

Quantitative attributes X_i impose an additional difficulty, as the attributes may assume a large or infinite number of values, thus the probability of a given value $X_i = x_i$ can be infinitely small *(Yang, 2009)*. In this situation the reliability of the estimation of $P(X_i = x_i \mid C = c)$ from the observed frequencies is not assured. There are two possible solutions to overcome this problem: *i*) estimate probabilities directly from the density function that gives the distribution of X_i over a specific class c. However, this density function must be estimated as it is usually unknown to real world data *(Yang, 2009)*; *ii*) discretization, a qualitative attribute X_i^* is formed for X_i where each value $X_i^* = x_i^*$ corresponds to an interval $(a_1, a_2]$ of X_i . Here, any $x_i \in (a_1, a_2]$ is replaced by x_i^* .

Thus, in order to correctly build the conditional probability tables the discretization of variables must be properly performed. Yang *(Yang, 2002)* describes several methods to perform that operation: Equal Width Discretization (EWD); Equal Frequency Discretization; Fuzzy Discretization, Entropy Minimization Discretization, Iterative Discretization, Proportional k-interval Discretization, Lazy Discretization, Non-disjoint Discretization, Weighted Proportional k-interval

Discretization. More recently the same author *(Yang, 2009)* reviewed some of the methods that are more often used for naïve Bayes classifiers.

Equal width discretization creates k intervals of equal width. Each interval has width Ω given through:

$$\Omega = \frac{max(x_i) - min(x_i)}{k} \tag{2.45}$$

$max(x_i)$ is the maximum value of X_i , $min(x_i)$ is the minimum value of X_i .

Equal frequency discretization divides the sorted values into k intervals so that each interval contains identical number of instances (N/k), where N is the total number of instances.

Both methods are frequently used for naïve Bayes classifiers due to their simplicity and reasonable good performance *(Yang, 2009).* However, in small samples the definition of a constant value for k may introduce some bias in the classification.

Entropy minimization discretization evaluates recursively the midpoint between each successive pair of the sorted values to identify the best cut value *(Yang, 2002).* The data is consecutively discretized between two intervals and the resulting class information entropy[34] is calculated. The cut point is selected as the one that presents the minimal entropy among all the candidates. The minimum description length (MDL) criterion is applied to provide the stop condition. This discretization might be effective at identifying decision boundaries in the one-attribute learning context, but in multi-attribute learning context (actual classification context), the resulting cut point can easily diverge *(Yang, 2009).*

Lazy discretization as a lazy approach postpones the discretization procedure to the classification time. When a test instance is presented, cut points for X_i are selected such that the value of X_i is in the middle of the corresponding interval. Lazy discretization creates only one interval for each variable and leaves the other region untouched. Probabilities are estimated from the training data, to build the conditional probability table and classify the query instance *(Hsu, 2003).* This method has high computational requirements which inhibit its utilization in large datasets *(Yang, 2009).*

---

[34] The entropy function of X is given by $H(X) = \sum_{i=1}^{n} p_i(x) \log\left(\frac{1}{p_i(x)}\right)$.

Discretization resulting in large interval frequency[35] tends to have low variance[36] in the same way discretization resulting in large interval number[37] tends to have lower bias[38] *(Yang, 2009)*. Some new techniques try to find a balance between these two concepts (interval frequency, interval number) in order to reduce the variance and bias. The proportional discretization method adjusts the number and size of discretized intervals to the number of training instances. Given N training instances the discretization is performed with $\sqrt{N}$ intervals, each one contains $\sqrt{N}$ instances:

$$I_f \times I_n = N \;;\; I_f = I_n \tag{2.46}$$

where I_f is the desired interval frequency and I_n the desired interval number. Values are sorted in ascending order and then I_n intervals of frequency I_f are created. Thus, if the dimension of training data set increases, the interval frequency and the number of intervals also increase which tend to have lower bias and lower variance *(Yang, 2009)*. This method has high potential to take advantage of training data sets with high dimension.

Fixed frequency discretization is an alternative approach to perform the discretization adjusted to the training data set dimension. An interval frequency I_f is defined, then the sorted values are grouped into intervals with that frequency. Each interval has approximately the same number of instances I_f including the possibility of having identical values (adjacent values). The number of intervals is directly proportional to the dimension of the training data set. This method is different from the equal frequency discretization since it does not define previously the interval number k. This difference allows the required flexibility to reduce variance and bias.

3. Semi naïve Bayes Methods

Naive Bayes makes the assumption that all the attributes X_i are conditionally independent given the value of class C. Although in many current situations this

---

[35] Interval frequency is the frequency of training instances in an interval formed by discretization.

[36] *Variance* describes the component of the classification error that results from random variation in the training data and from random behavior in the learning algorithm.

[37] Interval number is the total number of intervals formed by discretization.

[38] *Bias* describes the component of the classification error that results from systematic error of the learning algorithm.

assumption may be unrealistic. Some methods designated as semi naïve Bayesian methods, were developed to attenuate this attribute interdependence problem *(Zheng, 2005)*. The violation of the assumption of independence may not negatively affect the performance of naïve Bayes classifier *(Domingos, 1996) (Friedman, 1997) (Tsymbal, 2003),* however the performance of semi naive Bayesian methods suggest that the attenuation of the attribute independence can improve classification *(Zheng, 2005)*.

Semi naïve Bayesian methods can be divided in two main groups: *i*) methods that define a new set of attributes; *ii*) methods that define new interdependencies between attributes.

New set of Attributes

When two attributes are strongly correlated, naïve Bayes inference mechanism may overweigh their importance which increases the prediction bias.

Backward Sequential Elimination (BSE) considers the full set of attributes and successively eliminates the attribute whose elimination most improves accuracy, until there is no further accuracy improvement *(Zheng, 2005)*. BSE implements a naïve Bayes classifier with the remaining attributes.

Forward Sequential Selection (FSS) starts with an empty set of attributes and adds the attribute whose addition most improves accuracy. This process is an iterative process and stops when there is no further accuracy improvement. FSS also implements a naïve Bayes classifier with the final subset of attributes *(Langley, 1994)*.

Backward Sequential Elimination and Joining (BSEJ) creates new attributes (compound attributes) based on the original attributes and simultaneously evaluates the deletion of attributes. Values of the new compound attribute result from the Cartesian product of two original attributes' values *(Zheng, 2005)*. BSEJ implements a classifier that finds the class $c \in C$ through the following expression:

$$c = \underset{c_j}{argmax}\left(P(c_j)\prod_{r=1}^{h} P(jn_r \mid c_j)\prod_{i=1}^{q} P(x_i \mid c_j)\right) \tag{2.47}$$

where jn_r is the value of compound attribute $Jn_r \in Jn = \{Jn_1,...,Jn_h\}$, x_i is the value of original attribute $X_i \in \mathbf{X}$. BSEJ starts with the original attributes

representation and performs hill-climbing[39] search to find a new representation based on two operators *(Pazzani, 1996)*: *i*) replace a pair of original attributes by a new compound attribute that is the Cartesian product of the formers; *ii*) delete an attribute used by the attributes representation. The stop condition occurs when a change of representation does not result in an accuracy improvement (assessment is performed through leave one out cross validation[40]). Pazzani *(Pazzani, 1996)* stated that this approach has high potential to be applied in the enhancement of the naïve Bayes classifiers.

New Attributes Interdependencies

The methods that belong to this category intend to define new interdependencies among attributes. Here, the following algorithms are described: *i*) Tree-augmented naïve Bayesian (TAN); *ii*) Super Parent TAN (SP-TAN); *iii*) NBtree; *iv*) Lazy Bayesian Rules (LBR); *v*) Average One-Dependence Estimator (AODE).

The tree-augmented naïve Bayesian (TAN) is a variant of naïve Bayes, the structure $\mathcal{G}_T$ is composed of one class variable and several attributes. Similarly to naïve Bayes, the class variable has no parents and each attribute has the class variable as a parent. However, the TAN configuration (Figure 2.13) allows that one attribute can have at most one other attribute as a parent.

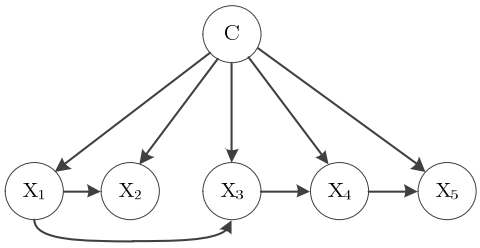


Figure 2.13 - Example of Tree Augmented Bayes network structure (Keogh, 1999).

[39] Iterative algorithm (optimization technique) that starts with an arbitrary solution to a problem, incrementally changes a single element of the solution in order to find a better solution. If better solution is found, an incremental change is made to the new solution, repeating until no further improvements can be found.

[40] In k-fold cross-validation, the original sample is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. In *leave one out cross validation* a single observation from the original sample is used as the validation data and the remaining observations as the training data.

The TAN algorithm must learn the proper structure (directed edges between the attributes) and may be described according to the algorithm detailed in Figure 2.14.

1. Compute mutual information $I_D(X_i;X_j \mid C)$ between each pair of attributes $i \neq j$

$$I_D(X_i;X_j \mid C) = \sum_{x_i,x_j,c} P(x_i,x_j,c)\log\frac{P(x_i,x_j \mid c)}{P(x_i \mid c)P(x_j \mid c)}$$

   This conditional mutual information function measures the information that X_j provides about the value of X_i given C.

2. Build a complete undirected graph in which the vertices are the attributes $X_i,...,X_j$
3. Build a maximum weighted spanning tree[41].
4. Transform the resulting undirected tree to a direct one by choosing a root variable and setting the directions of all edges to be outward from it.
5. Build a TAN model by adding a node C and adding an arc between C and each X_i.

Figure 2.14 - TAN algorithm (Friedman, 1997).

TAN algorithm classifies new instances through the inference mechanism that gives the class $c \in C$ according (*2.48*):

$$c = \underset{c_j}{argmax}\left(P(c_j)\prod_{i=1}^{p}P(x_i \mid c_j, Pa_{X_i})\right) \tag{2.48}$$

where Pa_{X_i} denotes the node that is the parent of X_i, $c_j \in C$ and x_i is the value of attribute $X_i \in \mathbf{X} = [X_1,...,X_p]$. As clearly depicted in Figure 2.13, the TAN structure weakens the independence assumption between attributes made by the naïve Bayes classifier.

Super Parent TAN (SP-TAN) was proposed by *(Keogh, 1999)* and it is a variant of TAN. It implements the same representation as TAN (Figure 2.13) but uses a different algorithm to find the new interdependencies between attributes. This method is based on three main concepts that are defined by Keogh as follows *(Keogh, 1999)*: *i*) *Orphan:* a node without a parent other than the class node is called an orphan; *ii*) *SuperParent:* if arcs are extended from X_i to every orphan, node X_i is a *SuperParent*; *iii*) *FavoriteChild*: if arcs are extended from node X_i to each orphan in

[41] Spanning tree is a subgraph that is a tree and connects all the vertices together.

turn, and test the effect on the predictive accuracy, the orphan that provides the best result is designated as *FavoriteChild* of X_i. The inference mechanism of SP-TAN is given by equation (2.48) and its algorithm can be described as follows:

1. Initialize network to naïve Bayes
2. Evaluate the current classifier
3. Evaluate each node as a *SuperParent*. Let X_{SP} be the *SuperParent* that originates a higher improvement in accuracy.
4. Define an arc from X_{SP} to each orphan. If the best arc improves accuracy: Then: keep it and return to step 2; Else: return the current classifier

Figure 2.15 - SP-TAN algorithm (Keogh, 1999).

NBtree was created by Kohavi *(Kohavi, 1996)* and combines decision trees with naïve Bayes classifier. The decision tree contains univariate splits (based on attributes value) as a regular decision trees, but the leaves contain naïve Bayes classifiers. Rather than the classical decision tree operation where the same class is predicted for all the instances that reach the leaf, the NBtree algorithm classifies the instances based on a naïve Bayes classifier built with the non-tested attributes. Thus, this hybrid approach takes advantage from the segmentation ability of decision tree and simultaneously from the evidence accumulation from multiple attributes provided by the naïve Bayes classifier *(Kohavi, 1996)*. A split is defined to be significant if the relative error reduction is greater than 5% and the splitting node has at least 30 instances *(Zheng, 2005)*.

The inference mechanism of NBtree is performed as follows:

$$c = \arg\max_{c_j} \left(P(c_j) \prod_{i=1}^{p-g} P(a_i \mid c_j, s) \right) \tag{2.49}$$

the parameter s is a specific value of $S = \{S_1,...,S_g\}$ that is the set of the g test attributes on the path to the leaf, a_i is a value of $A_i \in A = \{A_1,...,A_{p-g}\}$. that is the set of the remaining attributes, $c_j \in C$.

Lazy Bayesian Rules (LBR) is a lazy algorithm that generates a new Bayesian rule for each testing instance. The antecedent of a Bayesian rule is a set of conditions that are formed by the conjunction of several attribute/ value pairs. The selection of the attributes is guided by the specific testing instance. The consequent of Bayesian

rule is a local naïve Bayesian classifier created from the set of training instances that verify the antecedent rule. This classifier only uses those attributes that do not appear in the antecedent of the rule *(Zheng, 2005) (Zheng, 2000).*

As the antecedent of the Bayesian rule is composed by one or more than one attributes, the inference mechanism of Lazy Bayesian rules is also given by equation (2.49) where s is a value of $S = \{S_1,...,S_g\}$ that is the set of the tested attributes in the antecedent, a_i is a value of $A_i \in A = \{A_1,...,A_{p-g}\}$ the set of the remaining attributes, $c_j \in C$. Therefore, LBR can be seen as a branch of a tree generated by NBtree algorithm *(Zheng, 2005).* LBR generates a rule for each unseen instance while NBtree builds a single model considering all the examples in the training data.

Average One-Dependence Estimator (AODE) aggregates all predictions of a *one-dependence* classifiers. Thus, in AODE a one dependence classifier is built for each attribute in which the attribute is set to be the parent of all other attributes *(Webb, 2005).* In this way, in each one-dependence classifiers all attributes depend on the class and a single attribute. Given an instance to classify $\mathbf{x} = [x_1,...,x_p]$ and a parent's attribute value $X_i = x_i$:

$$P(c_j, \mathbf{x}) = P(c_j, x_i) P(\mathbf{x} \mid c_j, x_i) \tag{2.50}$$

This equation is valid for every x_i, which leads to:

$$P(c_j, \mathbf{x}) = \frac{\sum_{i:1 \le i \le p \wedge F(x_i) \ge m} P(c_j, x_i) P(\mathbf{x} \mid c_j, x_i)}{\left| \left\{ i : 1 \le i \le n \wedge F(x_i) \ge m \right\} \right|} \tag{2.51}$$

$F(x_i)$ is the frequency attribute-value x_i in the training sample. In order to assure the statistical significance of the obtained results, AODE averages only those models where the frequency of each attribute-value is larger than $m = 30$ *(Webb, 2005).* AODE algorithm classifies through:

$$c = \underset{c_j}{argmax} \left(\sum_{i:1 \le i \le p \wedge F(x_i) \ge m} P(c_j, x_i) \prod_{u=1}^{p} P(x_u \mid c_j, x_i) \right) \tag{2.52}$$

There are some variants of this algorithm such as: Weighted Average One Dependence Estimator (WAODE) that creates a weighted ensemble of one

dependence estimators; AODEsr that incorporates the lazy elimination of highly related attribute values at classification time *(Witten, 2011). (Zheng, 2006)*

4. Bayesian Classifiers Comparison

Zheng *(Zheng, 2005)* developed a very comprehensive study comparing the performances of the several Bayesian classifiers on 36 datasets of different dimensions as described in Table 2.8[42].

Nr.	Domain	I	A	C	Nr.	Domain	I	A	C
1	Adult	48842	14	2	19	Labor negotiations	57	16	2
2	Annealing	898	38	6	20	LED	1000	7	10
3	Balance scale	625	4	3	21	Letter recognition	20000	16	26
4	Breast cancer	699	9	2	22	Liver disorders	345	6	2
5	Chess	551	39	2	23	Lung cancer	32	56	3
6	Credit screening	690	15	2	24	Mfeat-mor	2000	6	10
7	Echocardiogram	131	6	2	25	New-thyroid	215	5	3
8	German	1000	20	2	26	Pen digits	10992	16	10
9	Glass ident.	214	9	3	27	Postop. patient	90	8	3
10	Heart	270	13	2	28	Primary tumor	339	17	22
11	Heart disease	303	13	2	29	Promoter Gene	106	57	2
12	Hepatitis	155	19	2	30	Segment	2310	19	7
13	Horse Colic	368	21	2	31	Sign	12546	8	3
14	House votes 84	435	16	2	32	Sonar Classification	208	60	2
15	Hungarian	294	13	2	33	Syncon	600	60	6
16	Hypotyroid	3163	25	2	34	Tic-Tac-Toe E.	958	9	2
17	Ionosphere	351	34	2	35	Vehicle	846	18	4
18	Iris classification	140	4	3	36	Wine recognition	178	13	3

I- instances; A – attributes; C – output classes

Table 2.8 - Testing datasets

Based on datasets described in Table 2.8, several semi naïve Bayes algorithms were compared as showed in Table 2.9:

	NB	AODE	NBtree	LBR	TAN	SP-TAN	BSEJ	BSE	FSS
Mean	0.220	0.206	0.214	0.211	0.219	0.212	0.213	0.218	0.241

Table 2.9 - Comparative of Bayesian classifiers – classification errors (Zheng, 2005)

[42] These datasets as well as their detailed description (attributes definition, class definition, etc) are available in http://archive.ics.uci.edu/ml/

Depending on the specific data set, it is possible to confirm that some of the semi naïve Bayes classifiers present lower classification errors than naïve Bayes classifier. This confirms that the attenuation of the attribute independence assumption can have a positive effect in the classification performance.

Webb *(Webb, 2005)* also performed a comparative study between Bayesian classifiers as presented in Table 2.10:

	Training			Classification		
	Time	μ	Space	Time	μ	Space
NB	$O(tn)$	3.41	$O(knv)$	$O(kn)$	2.92	$O(knv)$
TAN	$O(tn^2 + kn^2v^2 + n^2\log(n))$	8.60	$O(k(nv)^2)$	$O(kn)$	3.17	$O(knv^2)$
SP-TAN	$O(tkn^3)$	557.4	$O(tn + k(nv)^2)$	$O(kn)$	2.04	$O(knv^2)$
LBR	$O(tn)$	4.72	$O(tn)$	$O(tkn^2)$	85648	$O(tn)$
AODE	$O(tn^2)$	4.42	$O(k(nv)^2)$	$O(kn^2)$	22.1	$O(k(nv)^2)$

k: number of classes; n: number of attributes; v : average number of values for an attribute; t: number of training instances; μ : mean time (s).

Table 2.10 - Comparative of Bayesian classifiers (complexity, training time, testing time).

These results show that naïve Bayes has low complexity not only in the training phase but also in the testing phase. In fact, this classifier has lower training/testing time than the other Bayesian classifiers considered in the study.

These comparative studies show that in spite of some potential lack of accuracy, naïve Bayes is competitive with the other semi naïve Bayes classifiers. Additionally, it has the advantage of presenting lower training/testing time than the remaining classifiers.

In this thesis the implementation of the common representation of risk assessment tools based on naïve Bayes classifier seems appropriate. In fact, the selection of risk factors considered by each individual tool resulted from a statistical analysis process. This procedure usually starts with a large set of candidate risk factors, where the most relevant, typically not correlated, are selected. Therefore, the eventual violation of the attribute's independence is limited as this issue was already addressed in the statistical derivation of each individual risk assessment tool. Moreover, the proposed methodology addresses the potential lack of performance of naïve Bayes classifier through the implementation of an optimization procedure (genetic algorithm approach) that is carried out in the models' combination phase.

2.3 Models' Combination

As previously referred, the second step of the proposed methodology (Figure 1.2) is the combination of individual models, which aims to minimize some flaws of the current risk assessment tools, as it may: *i*) avoid the discarding of the available information originated by the previously developed tools; *ii*) allow the consideration of a higher number of risk factors; *iii*) avoid the choice of a "standard model". Additionally, the integration of individual classifiers can be very important for the improvement of the classification accuracy. A more accurate classifier might be obtained from a training dataset if several individual classifiers are trained and, after that, properly combined *(Tsymbal, 2003)*. Some methods to combine models are available in literature; however they can be organized according to two main categories: *i*) model output combination; *ii*) model parameter/data fusion[43].

2.3.1 Model Output Combination

According to various authors an ensemble of classifiers is often more accurate that any of the single classifiers in the ensemble *(Tsymbal, 2003) (Bauer, 1998).*

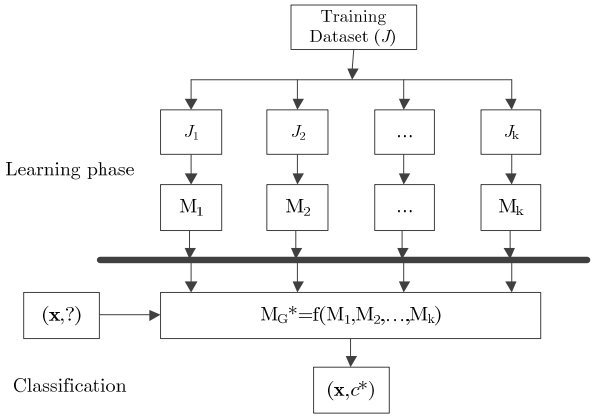


Figure 2.16 - Basic framework for an ensemble of classifiers (Tsymbal, 2003).

[43] Some authors designate model output combination as *model fusion*. Here, model parameter/data fusion refers to the direct combination of models' parameters rather than the combination of their outputs.

Figure 2.16 exemplifies this process, where the individual training data sets J_i are statistically independent, M_i denotes an individual model, M_G is the global model that is obtained through the integration of all M_i.

The integration procedure, i.e. the method to combine predictions of individual classifiers M_i, is essential to define the efficiency of combination. These methods can be classified in two main groups: *i*) static integration which applies the same method of combination for the entire data space; *ii*) dynamic integration that considers the characteristics of each specific instance to define the most proper combination procedure *(Tsymbal, 2003)*. Additionally integration methods can be separated in: *i*) Voting methods; *ii*) Selection methods.

Voting is the simplest integration method where the final result is based on votes of the individual models outputs. The output of each individual model is considered as a vote for a given class, the class with higher number of votes is identified as the final classification. Weighted voting is a more elaborate method, where the weight of each base classifier is assigned supported by the respective reliability *(Tsymbal, 2003)*. Thus, the importance of the vote (weight) is directly proportional to the reliability of each individual classifier and is applied to all tested instances. Dynamic voting states that the performance of each classifier can change according to the particular characteristics of each instance. Therefore the weights of the different classifiers are dynamically adjusted during the combination process *(Cordella, 1999)*.

According to Bauer *(Bauer, 1998)* voting classification methods can also be grouped in two types of algorithms: *i*) Bagging algorithms; *ii*) Boosting algorithms.

Bagging algorithm does not change the distribution of the training data set according to the performance of the individual classifiers. In this situation, the individual classifiers can be generated in parallel (Figure 2.17).

Input: training data set $J = \{(\mathbf{x}_1, c_1), ..., (\mathbf{x}_N, c_N)\}$; Learning Algorithm: $A(\cdot)$; integer: k (number of bootstrap samples)

for $i = 1$ to k {

J' =bootstrap sample extracted from J

$M_i = A(J')$ }

$$c = \underset{c_j \in C}{argmax} \sum_{i:M_i(\mathbf{x})=c_j}^{k} 1$$

Output: $c \in C$

Figure 2.17 - Bagging algorithm (Breiman, 1996).

Boosting algorithms generate the individual classifiers sequentially as the weight of each instance in the training data set is changed during the classification process. Several classifiers are built, each being trained on a data set where points which have been misclassified by the previous model have more weight *(Bauer, 1998)*.

Pan *(Pan, 2006)* also made a broad overview on voting methods, from which it is important to highlight two techniques: *i*) Bayesian model averaging; *ii*) Random forests.

Bayesian model averaging implements a weighted average of individual models, where the individual weights reflect how well the k individual models M_i fit the training dataset J.

$$\sum_{i=1}^{k} P(M_i \mid J) = 1 \tag{2.53}$$

The final classification is obtained through equation (2.54) *(Raftery, 2003) (Hoeting, 1999):*

$$P(C) = \sum_{i=1}^{k} P(c \mid M_i) P(M_i \mid J) \tag{2.54}$$

$P(c \mid M_i)$ is the probability of class c exclusively based on model M_i and $P(M_i \mid J)$ is the posterior probability of model M_i being correct given J. This probability (weight) is given by:

$$P(M_i \mid J) = \frac{P(J \mid M_i) P(M_i)}{\sum_{l=1}^{k} P(J \mid M_l) P(M_l)} \tag{2.55}$$

where $P(J \mid M_i)$ is the likelihood of model M_i and $P(M_i)$ is the prior probability that M_i is the true model. Although this technique has a high classification performance the calculation of the respective parameters may not be a straightforward process as explained in Hoeting *(Hoeting, 1999)*.

According to *(Breiman, 2001)*, *"A random forest is a classifier consisting of a collection of tree-structured classifiers* $\{M(\mathbf{x}, \Theta_k)\}$ *where* $\{\Theta_k\}$ *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input* $\mathbf{x}$*."*

Selection represents a different approach to the integration method as one of the individual classifiers is identified as the "best" model while the final classification is the result produced by it *(Tsymbal, 2003)*. According to Syed *(Syed, 2011)*, the best model is the smallest one that provides the best data description (*true model*).

The identification of "best" model can be done based on the minimization of an information criterion *(Zhang, 2009)*.

Akaike's information criterion (AIC) is given by:

$$AIC = -2log\gamma + 2\eta \tag{2.56}$$

the parameter γ is the maximized likelihood function of model M_i and η is the number of free parameters in the model. The second term of the AIC criterion represents a penalization factor for the model complexity. Thus, the most accurate model ("best" model) has the smallest value of AIC. Bayesian information criterion (BIC) differs from AIC only in the penalization factor, as presented in the equation (2.57); where N is the number of instances in the dataset. Zhang *(Zhang, 2009)* presented an overview of other information criteria that may be used to model selection.

$$BIC = -2log\gamma + \phi log\eta \tag{2.57}$$

Cross-Validation is a technique that can be applied to identify the individual classifier with the highest accuracy in the data space, which is partitioned into disjoint training dataset J and testing dataset O. An initial dataset $D = \{(\mathbf{x}_i, c_i), i = 1,...,N\}$ can be partitioned in two datasets such that $D = J \cup O$, with n instances in O and $(N-n)$ instances in J. The model M is fitted based on J which allows to obtain the class estimate $\hat{c}_O$ given O. There are $\binom{N}{n}$ possible partitions of data and the process can be repeated several times *(Syed, 2011)*. Some variants of cross validation may be implemented: *i*) Leave-one-out cross validation when $n = 1$; *ii*) Leave k-out cross validation where the size of O is k; *iii*) k-fold cross validation.

The method k-fold cross validation is frequently adopted since it is more efficient than the other variants of cross-validation. Here, the dataset $D = \{(\mathbf{x}_i, c_i), i = 1,...,N\}$

is divided into k partitions (folds) with approximately the same number of instances, such that:

$$D = \bigcup_{i=1}^{k} D_i \tag{2.58}$$

The model is trained in $(k-1)$ folds, the remaining fold is used for testing. This procedure is repeated k times so that each fold is used for testing one time. The selected model M_i will be the one that presents the lowest classification error.

Dynamic selection methods intend to select for each test instance the individual classifier that originates the most proper classification. Meta-level decision trees (MDT) algorithm is an example of this method *(Todorovsky, 2003)*. MDT has a structure that is identical to a regular decision tree, but rather than a class prediction this tree predicts which classifier should be used to classify a given instance. In the context of the selection based on information criteria, Shen *(Shen, 2004)* proposed an adaptive model selection that adjusts the penalization factor to the specific conditions of data.

Dynamic voting with selection combines the two approaches in order to increase the final classification accuracy *(Tsymbal, 2001)*. The errors originated by the base classifiers are estimated and the classifiers that present higher error values are discarded (selection). Then, a dynamic voting is applied to the remaining classifiers. A weight that is proportional to the respective estimated accuracy is assigned to the vote of each classifier.

2.3.2 Model Parameter/Data Fusion

This issue is less explored in bibliography than the models' output combination, although the direct combination of the parameters of the individual classifiers is potentially valuable.

Samsa et al. *(Samsa, 2005)* proposed a general regression strategy to combine risk factors of interest distributed across multiple datasets, providing a way to merge individual models into a multivariable risk model. In fact, many diseases have numerous risk factors, which are often studied in diverse cohorts with only a limited number of risk factors in each. The author proposes a method of combining univariate relative risks from diverse studies into a single multivariate model.

Steyerberg *et al. (Steyerberg, 2009)* presented a method to combine univariable regression results from the medical literature, with univariable and multivariable results from the individual patient dataset. They concluded that prognostic models may benefit from explicit incorporation of literature data.

Given the capabilities to deal with expert knowledge, Twardy *(Twardy, 2004) (Twardy, 2005)* proposed the use of Bayesian networks as a common approach to combine clinical expert knowledge with epidemiology models (Busselton, PROCAM) available in literature. The global model structure and parameters were derived from the published epidemiology models and supplemented by medical expertise.

Models' fusion approaches are being developed in several scientific domains, e.g. Logutov *(Logutov, 2005)* presented a work applied to ocean forecast[44]. The developed methodology designated by adaptive Bayesian model fusion intended to integrate multiple ocean models into a single ocean prediction system. This integration was based on the parameterization of individual forecast uncertainties *(Logutov, 2005)*.

The proposed combination methodology in this thesis fits in this category as it directly combines the parameters of individual models. Naïve Bayes classifier's structure is particularly indicated to allow the direct integration of individual models' parameters. This approach can be very flexible since it incorporates different individual contributions into a global model that can be adjusted to the specific test conditions.

Additionally, the resulting global model preserves the main characteristics of individual Bayesian models such as: *i*) interpretability of the model; *ii*) ability to deal with missing risk factors.

2.3.3 Optimization

Models' combination is responsible for the creation of the global model that assesses the risk of occurrence of a CVD event. However, its parameters can be adjusted in order to increase the performance of the model.

In this context, optimization methods must also be addressed, since they contribute significantly to the accuracy's improvement of the proposed combination scheme.

[44] Several parameters may be forecasted e.g. temperature, sea surface height, sea surface salinity, etc.

An optimization problem can be formulated as:

$$\begin{aligned} & \min_{\mathrm{x}\in\mathrm{R}^n} f(\mathrm{x}) \\ & restricted\ to: \\ & \quad l_{ri}(\mathrm{x}) = 0, \quad i = 1,...,m \\ & \quad l_{rj}(\mathrm{x}) \geq 0, \quad j = m+1,...,t_c \end{aligned} \tag{2.59}$$

where $f(\mathrm{x})$ is designated by the objective function; $l_{ri}(\mathrm{x}) = 0$ represent the m equality constraints and $l_{rj}(\mathrm{x}) \geq 0$ are the $(t_c - m)$ inequality constraints.

There are multiple methods to solve an optimization problem. Their selection depends on several criteria such as: type of objective function (e.g. differentiable or not differentiable), variables' type (Boolean, discrete, continuous, etc.), type of considered constraints and type of solution (local minimum, global minimum).

It is possible to identify two main categories of iterative methods for optimization: *i*) classical optimization; *ii*) heuristic optimization.

Classical optimization methods are characterized by an analytical condition or gradient-based approach where an individual solution is found in a single iteration (candidate solution). This solution is refined until some criterion is met or a certain number of iterations have been performed *(Peddersen, 2010) (Anile, 2005)*. Descendent methods, Quasi-Newton method, Levenberg-Marquardt method are examples of classical optimization methods *(Fletcher, 1999) (Michalewicz, 2004)*.

According to Gilli *(Gilli, 2008)*, heuristic optimization methods should be able to provide high quality approximations (stochastic nature but controlled) to the global optimum. These methods should not be too sensitive to some changes in the search space or in the algorithm's parameters. Finally, they should be easily implemented to several instances of the considered problem. Genetic algorithms, ant colonies, differential evolution, particle swarm optimization are examples of population-based heuristic optimization methods *(Michalewicz, 2004) (Lee, 2008)*.

In this thesis, the application of genetic algorithms seems appropriate as they can be applied to both constrained and unconstrained optimization problems and present a very competitive performance when compared with iterative classical methods *(Anile, 2005) (Luong, 2003)*. In fact, the objective function $f(\mathrm{x})$ adopted to the adjustment of the parameters (probabilities) of the Bayesian global model is not differentiable which obstructs the utilization of gradient-based optimization methods. Additionally, genetic algorithms can deal with constraints which impose that the

adjustment of parameters should be done neighboring the original values (assure the model's clinical interpretability).

Genetic Algorithms

Genetic algorithms are inspired by the evolutionist theory, which establishes that in nature, weak and unfit species within their environment are faced with extinction by natural selection. Thus, the strong ones have greater opportunity to pass their genes to future generations via reproduction *(Konak, 2006).*

Genetic algorithms operate with a set of candidate solutions (*population*). A solution (also designated as individual or chromosome) is usually a vector composed by a set of discrete units called *genes*. The initial population is randomly established and its size, that is constant during the evolutionary search, is usually higher or equal to the number of the parameters to optimize. Genetic algorithms iteratively modify a population of solutions in a sequential way (Figure 2.18). As the search evolves, the population has fitter and fitter solutions, and eventually converges, meaning that it is dominated by a set of similar solutions *(Eiben, 2003).*

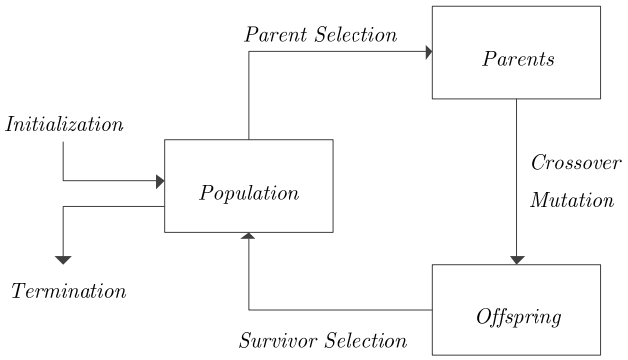


Figure 2.18 - General scheme of a genetic algorithm.

Each iteration cycle of the GA is composed of several steps. An evaluation function must be defined to assign a quality measure to each solution. Then a *Parent Selection* mechanism must be implemented to distinguish among individuals based on their quality, in particular to allow the better individuals to become parents of the next generation. This parent selection process is typically probabilistic, thus high quality individuals have a higher chance of becoming parents than those individuals with low quality.

After parent selection a new set of individuals should be created (*Offspring*). This is the role of the stochastic variation operators (*Crossover*, *Mutation*) that

create new individuals from the old ones. The operator designated by *Crossover*[45] is binary, since it merges information from two individuals (parents) into one or two offspring individuals. *Mutation* is a unary operator that is applied to solutions resulting from crossover operation and delivers a modified mutant (child/offspring) of itself. This operator is also stochastic since its output depends on the outcome of a series of random choices.

The *Survivor Selection* is responsible for the selection of next generation individuals based on the individuals of the current *Population* and on the derived offspring.

The population evolves, the cycle iterates, until *Termination*[46] condition is reached. There are some conditions that are often used such as: *i*) fitness limit; *ii*) time limit; *iii*) maximum number of new generations; *iv*) fitness improvements remain under a threshold value during a given period of time.

It is important to detail each one of the main steps that compose the genetic algorithm operation (Figure 2.18).

1. Representation of individuals

The first step of GA's application is to define a proper representation of the candidate solutions. This decision can have a main impact in the optimization performance of the GA.

Binary representation of individuals consists of a string of binary digits. For problems involving Boolean variables this is the natural representation to be adopted. Binary representations can also be applied to encode non-binary information[47] although in these cases the optimization results obtained can be biased.

Integer representation is often more appropriate when the genes[48] can assume a discrete set of values. This representation can be: *i*) unrestricted, when the genes may represent any integer value; *ii*) restricted to a specific set (e.g., $\{0,1,2,3\}$ representing $\{North, East, South, West\}$) *(Eiben, 2003).*

---

[45] Crossover operator is also designated by Recombination.

[46] Termination condition is also designated by stopping condition.

[47] Gray coding may be used in the conversion binary-integer, as it assures the same Hamming distance (one) between consecutive integers.

[48] An individual is made of discrete units called genes.

Floating-point representation is made through a string of real values that is applied when the values to be represented come from a continuous distribution. In this case the individual is a vector $\boldsymbol{a}=[a_1,...,a_k]$; $a_i \in \mathbb{R}$, k is the number of genes of the individual.

A specific type of representation is designated by permutation and it is applied for problems when the value of a gene must be unique in the individual, e.g. find the order that a sequence of events should occur. These representations can be implemented through an encoded permutation of a set of integers, e.g. the j element of the representation denotes the event that happens in j ($[A,B,C,D]$ with the permutation [3,1,2,4] originates the solution $[C,A,B,D]$) *(Eiben, 2003)*.

2. Parent Selection

A probability distribution to define the likelihood of each individual in the population to be selected for reproduction must be implemented. There are two distributions that should be referred: *i*) fitness proportional selection; *ii*) ranking selection.

Fitness proportional selection is based on the probability (2.60) that an individual i is selected among the μ individuals that integrate the population, i.e. the selection probability depends on the absolute fitness of the individual when compared to the absolute fitness value of the rest of population.

$$P_{sel}(i) = \frac{f_i}{\sum_{j=1}^{\mu} f_j} \tag{2.60}$$

Table 2.11 shows an example of this type of selection *(Eiben, 2003)*:

String n°	Initial Population*	x value	Fitness $f(x)=x^2$	$P_{sel}(i)$	Expected Number
1	01101	13	169	0.14	0.58
2	11000	24	576	0.49	1.97
3	01000	8	64	0.06	0.22
4	10011	19	361	0.31	1.23

Maximize fitness function $f(x)=x^2$ *; *5 bit binary encoding of the x value ;* $\sum_{j=1}^{\mu} f_j = 1170$ *;* $\overline{f} = 293$

Table 2.11 - Fitness proportional selection example.

As the number of parents is constant (population's size (μ)), the expected number of copies of each individual is given by $f_i/\bar{f}$. This fitness proportional selection may originate two different problems: *i*) premature convergence, when there are individuals that are much better than the rest of population; *ii*) no selection pressure, when the fitness values of individuals are very close. In the latter the selection of the best individuals may have little impact on the improvement of performance[49].

Ranking selection is an alternative method to the parent selection. It sorts the individuals based on their fitness and then defines probabilities for the individuals according to their rank. This mapping between rank position and selection probability may be implemented in different ways (e.g. Table 2.12).

Individual	Fitness	$P_{sel}(i)$	Rank	$P_{sel}(i)^*$
A	1	0.1	0	0
B	5	0.5	2	0.67
C	4	0.4	1	0.33

**one possible probabilities' definition considering the rank position*

Table 2.12 – Ranking selection example.

The two described methods (fitness/rank) define the likelihood of each individual being selected for reproduction. Additionally, algorithms to implement the selection of parents must be considered.

The roulette wheel algorithm (Figure 2.19) is the simplest method to select the individuals based on their selection probabilities. It assumes that there is an order (random or ranked) over the population from 1 to μ where a set of values $\mathbf{b} = [b_1 ... b_\mu]$ is calculated based on:

$$b_i = \sum_{j=1}^{i} P_{sel}(j) \tag{2.61}$$

where $P_{sel}(j)$ is given by (2.60) and $b_\mu = 1$. The value r is randomly picked from the interval $[0,1]$, and the selected parent corresponds to the first value of $\mathbf{b}$ that is higher than r *(Eiben, 2003)*[50]. This process is repeated μ times (obtain μ parents).

[49] A more detailed discussion on this topic can be found in *(Eiben, 2003)*

[50] This author states that conceptually this method is the same as spinning a one armed roulette wheel where the sizes of the holes reflect the selection probabilities.

Population's size (parents) μ ; $\mathbf{b} = [b_1,...,b_\mu]$ such that $b_i = \sum_{j=1}^{i} P_{sel}(j)$ where the $P_{sel}(j)$ is defined by fitness proportional or ranking selection:

```
BEGIN
    set current_member=1:
    WHILE (current_member ≤ μ) DO
        pick a random value r from [0,1]; set i = 1;
        WHILE (b_i ≤ r) DO
            set i = i + 1; END
        set parents_pool[current_member]=parents[i];
        set current_member= current_member+1;
    END
END
```

Figure 2.19 - Roulette wheel algorithm (Eiben, 2003).

Stochastic universal sampling is an evolution of the roulette wheel algorithm that produces a better sample of the required distribution. It diverges from the previous algorithm in the initialization of r that is made in the interval $[0,1/\mu]$ along with the subsequent update of $r = r + 1/\mu$. This enhanced definition of r assures that the parent selection reflects more accurately than the roulette wheel algorithm the estimated likelihoods (fitness/rank) of individuals to be selected for reproduction.

```
BEGIN
    set current_member=1;
    i=1; pick a random number r from [0,1/μ];
    WHILE (current_member ≤ μ) DO
        WHILE (r ≤ a[i]) DO
            set parents_pool[current_member]=parents[i];
            set r = r + 1/μ;
            set current_member= current_member+1;
        END
        set i = i + 1;
    END
END
```

Figure 2.20 – Stochastic universal sampling algorithm (Eiben, 2003).

Alternatively, Tournament selection is an operator that does not require any global knowledge of the entire population. This algorithm operates based on an

ordering relation (e.g. fitness value) that can rank r randomly picked individuals from the population. The selection of an individual as the winner of the tournament depends of several factors: *i*) its rank in the population; *ii*) the tournament size (r elements); *iii*) if the individuals are chosen with or without replacement; *iv*) the probability that the tournament's member with the highest fitness/rank is selected *(Bäck, 1995) (Eiben, 2003)*.

3. Variation Operators

Variation operators are responsible for the generation of the offspring and can be grouped in two main types: *i*) crossover, where a child is created from the combination of two or more parent solutions, *ii*) mutation, when one the gene of a solution is modified to generate one child.

The stochastic application of both variation operators is controlled by parameters of the algorithm, designated as crossover rate[51] and mutation rate[52]. The application of the variation operators allows the creation of a new set of individuals (offspring) composed by a combination of information from the current population that generates promising solutions.

Crossover operators depend on the adopted representation as well as on the specific properties of the encoded information. For binary as well as for integer individuals, one-point crossover is usually adopted and it operates as described in Figure 2.21

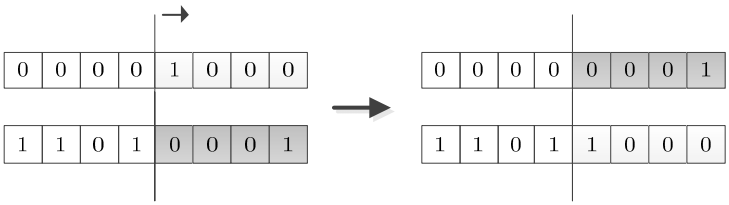


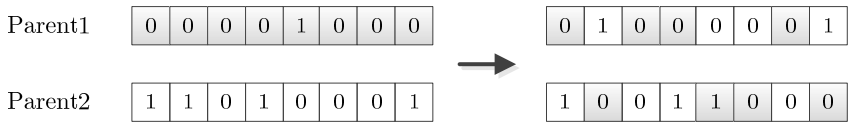
Figure 2.21 – One-point crossover.

An extension of one-point crossover is the N-point crossover, where $N+1$ segments are switched to create the offspring.

[51] Probability of applying crossover to a pair of parents.

[52] Probability of mutating a specific gene.

Uniform crossover is a generalization of N-point crossover and is based on a set of values generated from a uniform distribution. This set, that has the same dimension as the number of genes, is responsible for the selection of the parent, e.g. given two parents (parent 1 and parent 2) if values are below a given threshold then the child inherits the value from parent 1 otherwise the child inherits from parent 2. The second child is obtained using the inverse mapping (Figure 2.22).



[0.4, 0.6, 0.1, 0.3, 0.7, 0.8, 0.4, 0.6] (if value < 0.5 then parent1 otherwise parent2)

Figure 2.22 – Uniform crossover.

Crossover operators for floating-point representations can be of two different types: *i*) discrete recombination, similar to the crossover operators for binary and integer individuals; *ii*) arithmetic recombination, where in each gene of the offspring a new value that lies between those of the parents ($z_i = \alpha p_i^1 + (1-\alpha)p_i^2;\ \alpha \in [0,1]$) is created; z_i represents the new value of the gene i, p_i^1, p_i^2 represent respectively the values of gene i in parent 1 $\mathbf{p}^1 = [p_1^1 .... p_k^1]$ and parent 2 $\mathbf{p}^2 = [p_1^2 .... p_k^2]$.

The simple recombination implements a recombination from a specific point λ, such as: $\mathbf{ch}^1 = [p_1^1, ..., p_\lambda^1,\ (\alpha p_{\lambda+1}^2 + (1-\alpha)p_{\lambda+1}^1), ...,\ (\alpha p_k^2 + (1-\alpha)p_k^1)]$, where $\mathbf{ch}^1$ represents the child 1 and k is the total number of genes of the individual. Child 2 is defined similarly as child 1: $\mathbf{ch}^2 = [p_1^2, .., p_\lambda^2,\ (\alpha p_{\lambda+1}^1 + (1-\alpha)p_{\lambda+1}^2), ..,\ (\alpha p_k^1 + (1-\alpha)p_k^2)]$.

The single arithmetic recombination implements a recombination only in the randomly selected position λ: $\mathbf{ch}^1 = [p_1^1, ..., p_{\lambda-1}^1,\ (\alpha p_\lambda^2 + (1-\alpha)p_\lambda^1),\ p_{\lambda+1}^1, ..., p_k^1]$[53].

The whole arithmetic recombination is obtained for each offspring gene ($i \le k$) through the weighted sum of the respective parent genes' values: $\mathbf{ch}^1 = \alpha p_i^1 + (1-\alpha)p_i^2$ and $\mathbf{ch}^2 = \alpha p_i^2 + (1-\alpha)p_i^1$.

There are several crossover operators specific for individuals represented through permutations: *i*) partially mapped crossover; *ii*) edge crossover; *iii*) order crossover; *iv*) cycle crossover. These methods aim to transmit the information contained in

[53] $child_2$ is obtained swapping $\mathbf{p}^1 = [p_1^1 .... p_k^1]$ and $\mathbf{p}^2 = [p_1^2 .... p_k^2]$.

parents especially the information they hold in common *(Eiben, 2003)*. The detailed description of these recombination algorithms is out this thesis' scope.

Mutation operators are also defined according to the specific characteristics of the individuals' representation. For binary encoding, the mutation operator considers each gene individually. The values of different genes may be flipped according to a probability value (bitwise mutation rate).

Figure 2.23 - Binary encoding mutation operator (example).

Random resetting mutation is applied to integer representations of individuals. This operator is similar to the previous one (Figure 2.23). A new value for each gene (according to a given probability) may be chosen at random from the set of permissible values. Creep mutation is an alternative mutation operator that adds a small value (positive/negative) to each gene (according to a given probability). This method is more likely to generate small changes than large ones *(Eiben, 2003)*.

The mutation operator for floating-point representations is based on a continuous distribution and can assume two different types: *i*) uniform mutation, where values of the child's genes are drawn randomly from a specified range of values; *ii*) nonuniform mutation, similar to creep mutation, in this situation the new value is obtained through the addition of an amount randomly drawn from a Gaussian distribution with the mean value set to zero and standard deviation defined by the user.

The mutation operators for permutation are based on changes of the genes' values but restricted to the original ones (parent's values). Swap mutation randomly picks two genes in the individuals and swaps their values (Figure 2.24).

Figure 2.24 – Swap mutation.

Insert mutation selects two genes at random and moves one of them as described in Figure 2.25.

Figure 2.25 – Insert mutation.

Scramble mutation is based on the selection of a subset of genes and then scrambles their positions (Figure 2.26).

Figure 2.26 – Scramble mutation.

Inversion mutation randomly selects two genes and reverses the order in which the values appear between those positions (Figure 2.27).

Figure 2.27 – Inversion mutation.

4. Survivor Selection

This step is responsible for the selection of the individuals of the next generation based on the individuals of the current generation and on the respective offspring. There are some replacement strategies: *i*) age based replacement; *ii*) fitness based replacement.

Age based replacement operates depending on the number of cycles that an individual exists. It can be simply implemented considering that one cycle is the duration of all the individuals, which forces the replacement of all parents by the entire offspring (the population size remains constant) in each cycle. Here, it is important to refer the elitism mechanism, which assures that the current fittest member is not discarded during the genetic algorithm operation. If the individual is selected to be replaced but none of the offspring individuals have a higher fitness value than that individual, then it is maintained in the population and one of the offspring individuals is discarded. Alternatively, replacement can be based on the fitness value. The individuals from the current population as well as from the respective offspring are ordered and the best μ individuals are selected to integrate the next generation (*Population*) as depicted in Figure 2.18.

The cycle iterates until a termination condition is reached, as mentioned there are some termination conditions that are often used (e.g. fitness limit, time limit, fitness improvements below a threshold during a given time, etc.).

In this thesis all of these concepts were considered to obtain an optimization algorithm able to improve the performance of global model.

In fact, the proper selection of the genetic algorithm's parameters (parents' selection algorithm, variation operators, termination condition, etc.) is important to maximize its efficiency. However, the tuning of a genetic algorithm's parameters can be very challenging and time consuming.

2.3.4 Missing Information

One of the main objectives of this thesis is to increase the ability of the risk assessment tools to deal with missing information (missing risk factors). Actually, missing risk factors is a very serious and frequent problem that must be circumvented in the physicians' daily activity. As already referred, *"... information on patients such as demographic data, medical history, treatments, test results, and family structure is often unavailable when a doctor greatly needs" (Khanna, 2005).*

First of all, it is important to identify the mechanisms that originate the missing risk factors: *i*) Missing completely at random (MCAR); *ii*) Missing at random (MAR); *iii*) Not missing at random (NMAR) *(Steyerberg, 2009).*

The subjects with missing information in the MCAR mechanism are representative of the population with complete data. Missing data can be caused by random factors (handling error, breakdown of equipment, administrative error). MAR situation occurs when the probability of a missing predictor is independent of the risk factor itself, but depends on the observed values of other variables, e.g. age, missing values increase in older patients. NMAR mechanism happens when the missing values depend on the predictor itself (e.g., personal data income, sexual orientation, etc.) or they are based on other predictors that are not observed (e.g. body mass index's value/obesity condition, etc.) *(Graham, 2003).*

Depending on the type of missing data there are different methods that can be applied to deal with this situation. The simplest method is *listwise deletion* or *complete case analysis (Wayman, 2003) (Horton, 2007)*, where the instances with missing data are omitted. This method can be adopted in statistical inference applications if the instances with missing data represent less than 5% of the total number of cases. Otherwise, it represents a significant loss of statistical power and can originate an estimation bias. This is not a valid method to be applicable to the individual patient risk prediction.

Mean substitution is another method that replaces the missing value of a variable with the mean value of that variable. This simple imputation method has a serious drawback as the variance of the respective risk factor is artificially reduced. Additionally the relationships with other variables may also be influenced. Subgroup mean imputation is a variant of the mean substitution, where the missing risk factor's value is imputed with a subgroup mean value. The subgroup is created from the derivation set based on a subset of variables (e.g. based on clustering techniques) *(Janssen, 2009)*.

Other methods that are statistically supported do not concentrate just on identifying a replacement for a missing value but also on preserving the relationships between the several risk factors. Multiple imputation (MI) is often referred to as a method that produces an accurate prediction of missing values *(Wayman, 2003) (Horton, 2007) (Janssen, 2009)*.

In MI missing values of a specific variable are predicted using other variables that also belong to the dataset. Several regression models are created producing the respective imputed data sets. The overall analysis is obtained considering the standard statistical analysis performed in each imputed data set. Thus, multiple imputation can be systematized in three steps: *i*) creation of imputed data sets; *ii*) statistical analysis of each one; *iii*) combination of the statistical analysis results *(Wayman, 2003)*. The global mean can be calculated by averaging the individual means:

$$\overline{x}_i = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i \tag{2.62}$$

the parameter n is the number of imputed data sets and $\hat{x}_i$ is the estimate of each $x_i \in \mathbf{x}$ individual mean. The total variance can be obtained through the following expression:

$$V = \overline{W} + (1+\frac{1}{n}) \times U \tag{2.63}$$

where:

$$\overline{W} = \frac{1}{n}\sum_{i=1}^{n} W_i \quad ; \quad U = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{x}_i - \overline{x}_i)^2 \tag{2.64}$$

the variable $\overline{W}$ measures the natural variability of data (W_i is the variance of individual estimates) and U measures the uncertainty due to imputation.

As stated in *(Janssen, 2009)*, multiple imputation is straightforward and feasible when analyzing a whole dataset. The application of this technique to an individual patient is more complex. Firstly the specific patient must be added to a proper dataset and then the multiple imputation is performed.

Horton *(Horton, 2007)* identifies other methods to deal with missing risk factors (likelihood based approaches, weighting methods, etc.) though these methods are not frequently used *(Burton, 2004)*.

The probabilistic classifiers, namely the naïve Bayes inference mechanism, deal with missing information in a different perspective, as it does not need a specific value's imputation. Actually, the naïve Bayes inference mechanism prevents the degradation of the predictive performance of the model disabling the influence of the missing risk factor. This particular feature of Bayesian inference mechanism will be explored and compared with some of the mentioned imputation methods in order to reach a conclusion on the reliability of the prediction in the presence of missing risk factors.

2.4 Grouping of Patients

2.4.1 Dimensionality Reduction

The dimensionality reduction (DR), aiming for the creation of a low dimensional representation of a high dimensional data sample while preserving most of the intrinsic information[54] contained in the original data, can be very useful in several applications. In fact, the reduction of dimensionality is important in many domains since it reduces the curse of dimensionality[55] and other undesired properties of high-dimensional spaces *(Sugiyama, 2010)*. Alternatively, the main goal of dimensionality

[54] The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of data *(Sugiyama, 2010)*.

[55] Curse of dimensionality refers to the fact that in the absence of simplifying assumptions the number of data samples required to estimate a function to a given accuracy grows exponentially with the number of dimensions *(Lee, 2007)*.

reduction may be simply stated as: "*a large set of parameters or features must be summarized into a smaller set, with no or less redundancy*[56]" *(Lee, 2007)*.

The dimensionality reduction can be formalized as adopted by Fodor *(Fodor, 2002)*: given a p dimensional data vector $\mathbf{x} = [x_1 ... x_p]^T$, a lower dimensional representation $\mathbf{y} = [y_1 ... y_q]^T$ should be found with $q \leq p$ and $\mathbf{y} \not\subset \mathbf{x}$ such that it captures the content in the original data according to some criterion.

Lee *(Lee, 2007)* identified several possible qualifications of dimensionality reduction methods. From this set, the following possible classifications should be highlighted:

- Hard *vs.* soft dimensionality reduction, where the ratio between the initial and the reduced dimensions is applied to distinguish the two categories;
- Supervised *vs.* unsupervised approaches. Rather than unsupervised approaches that estimate the reduced dimensions exclusively based on the input data, supervised methods calculate the projections considering both input and output data;
- Linear vs. nonlinear, linear techniques assure that each one of the $q \leq p$ elements of the new data space is a linear combination of the original variables. Nonlinear techniques are able to deal with complex nonlinear data *(Maaten, 2009)*;
- Continuous *vs.* discrete model. A discrete model implements a finite set of interconnected points between the two dimensional spaces;
- Layered *vs.* standalone embeddings. The methods that produce standalone embeddings must compute all the parameters for a required dimensionality reduction every time that the target dimensionality changes;
- The type of criterion to be optimized. For instance distance preservation, where the pairwise distances measured between data points in the reduced space should be as similar as possible to the ones verified in the original space.

[56] Redundancy means that parameters or features that could characterize an individual (unit) are not independent from each other *(Lee, 2007)*.

However, the most frequently adopted classification of dimensionality reduction methods is: *i*) linear methods; *ii*) non-linear methods *(Maaten, 2009) (Sugiyama, 2010) (Fodor, 2002) (Lee, 2007).*

Linear Methods

The methods that integrate this category assume that each one of the $q \leq p$ components y_i, $i = 1,...,q$ of the new instance $\mathbf{y}$ is a linear combination of the original variables:

$$y_i = w_{i,1}x_1 + ... + w_{i,p}x_p, \quad i = 1,...,q \tag{2.65}$$

considering the matrices operation:

$$\boldsymbol{Y}_{q\times N} = \boldsymbol{W}_{q\times p}\boldsymbol{X}_{p\times N} \tag{2.66}$$

where p is the dimension of original data, q denotes the dimension of the lower dimensional representation and N is the number of instances. Among the linear methods it is possible to identify the Principal Component Analysis (PCA), Independent Component Analysis, Factor Analysis *(Maaten, 2009) (Fodor, 2002).*

The PCA is possibly the most common linear technique applied in dimensionality reduction. It reduces the original dimension of data by finding a few orthogonal linear combinations with the largest variance. The first principal component $y_1 = \mathbf{x}^T\mathbf{w}_1$ is the linear combination with the largest variance such as:

$$\mathbf{w}_1 = \textit{argmax}_{\|\mathbf{w}=1\|}\mathrm{Var}\{\mathbf{x}^T\mathrm{w}\} \tag{2.67}$$

with $\mathbf{w}_i = [w_{i,1}....w_{i,p}]^T$. The second principal component is the linear combination with the second largest variance and the same reasoning is applied to the remaining components. There are as many principal components as the number of original variables. However, for the majority of the applications, the first components explain most of the variance. This allows the elimination of the remaining principal components with minimal loss of information *(Fodor, 2002).*

Assuming that the covariance matrix $\boldsymbol{\Sigma}_{p\times p}$ of $\boldsymbol{X}_{p\times N}$ [57] is decomposed as:

---

[57] The calculation of the covariance matrix, first step of PCA, requires the previous standardization of observations $x_{i,j}$ by $(x_{i,j} - \hat{\mu}_i)\big/\hat{\sigma}_i$ where $\hat{\mu}_i = \sum_{j=1}^{N} x_{i,j}\big/N$ and $\hat{\sigma}_i = \sqrt{\sum_{j=1}^{N}(x_{i,j} - \mu_i)^2\big/N}$.

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{2.68}$$

where $\mathbf{\Lambda} = diag\,(\lambda_1,...,\lambda_p)$ is the diagonal matrix of the ordered eigenvalues $\lambda_1 \leq ... \leq \lambda_p$ and $\mathbf{U}$ is a $p \times p$ orthogonal matrix containing the eigenvectors[58]. The final data (dimensionality reduction) is obtained through the following expression:

$$\mathbf{Y}_{p\times N} = \mathbf{U}^T_{p\times p}\mathbf{X}_{p\times N} \tag{2.69}$$

The q first elements (rows) of $\mathbf{Y}$ are considered while the remaining $(p-q)$ are discarded, which permits a reduction from p to q dimensions.

Non Linear Methods

Here, it is not possible to determine a linear transformation weight matrix $\mathbf{W}$ between dimensional spaces p and $q \leq p$. Maaten *(Maaten, 2009)* who presents a very comprehensive overview on non-linear methods defines three main categories: *i*) global techniques; *ii*) local techniques; *iii*) global alignment of linear models.

Global techniques for dimensionality reduction attempt to preserve the global properties of data *(Maaten, 2009)*. Multidimensional scaling aims to retain the pairwise distances between the original and the respective reduced data space. The target is the minimization of the error between the pairwise distances in the high dimensional and low dimensional representation. The following equation shows a possible criterion to be minimized:

$$\phi = \sum (\left\|\mathbf{x}_i - \mathbf{x}_j\right\| - \left\|\mathbf{y}_i - \mathbf{y}_j\right\|)^2 \tag{2.70}$$

In some specific situations Euclidean distance may originate biased results. Isomap intends to circumvent this difficulty. It considers the geodesic[59] (curvilinear) distances between data points. In this case, geodesic distances between the data instances $\mathbf{x}_i$, $i = 1,...,N$ are computed by constructing a neighborhood graph in which every point is connected with its k nearest neighbors in the dataset. The

---

[58] An eigenvector $\mathbf{u}$ of matrix $\mathbf{A}$ can be defined as the solution of equation $\mathbf{Au} = \lambda\mathbf{u}$ where λ is the eigenvalue (scalar) associated to the eigenvector $\mathbf{u}$.

[59] Geodesic distance is the distance between two points measured over the manifold, which is an abstract mathematical space in which every point has a neighborhood. This neighborhood resembles the geometry spaces described by Euclidean geometry *(Ribeiro, 2008)*.

shortest path between two points is a good estimate of the geodesic distance between two points. Global techniques category comprises other methods such as: maximum variance unfolding, diffusion maps, neural networks *(Maaten, 2009).*

Local nonlinear techniques for dimensionality reduction intend to preserve properties of small neighborhoods around the data instances (consider the nearest neighbors). Local linear embedding, Laplacian eigenmaps, Hessian local linear embedding are examples of local techniques.

Global alignment of linear models techniques combines global and local techniques. According to Maaten *(Maaten, 2009)* they compute a number of local linear models and perform a global alignment of these linear models. Locally linear coordination (LLC) and manifold charting are included in this category.

This brief overview of dimensionality reduction techniques intends to give a global perspective of the main techniques that are applied to reduce the dimension of data spaces. However, this is a very extensive topic whose detailed exploitation is beyond the scope of this thesis.

In this work, taking advantage of the specificities of CVD risk assessment, a specific method is applied to implement the dimensionality reduction. This step is particularly important in the development of the strategy described in Section 3.6.

2.4.2 Clustering

Clustering techniques have a very relevant role in unsupervised learning, where the goal is to find a suitable representation of underlying distribution of the unlabeled data. According to Fung *(Fung, 2001)* clustering is defined as: "*Clusters should reflect some mechanism at work in the domain from which instances or data points are drawn, that causes some instances to bear a stronger resemblance to one another than they do to the remaining instances.*

Thus, clustering techniques group data objects into clusters such that: $\Upsilon \in R^{N \times p}$ represent a set of N instances $\mathbf{x}_i \in R^p$, the goal is to partition Υ into K groups $G = \{G_1,...,G_K\}$ where data that belong to the same group are more similar than data that belong to the other groups.

This capability to find similarities between data objects based on the underlying structure of the dataset can be very useful in the context of the CVD risk assessment. The discovery of similarities between patients (creation of groups of patients) can be

important to identify the model that presents the best performance within a specific group of patients.

The concepts of similarity and dissimilarity are defined based on the resemblance between data instances, which can be assessed based on different distance metrics that are selected according to the type of data. Table 2.4 can be expanded through the integration of the main distance metrics organized by the type of attribute (Table 2.13):

Interval Scaled Attributes	Euclidean	$d(\mathbf{u},\mathbf{v}) = \sqrt{\sum_{i=1}^{p}(u_i - v_i)^2}$
	Minkowsky	$d(\mathbf{u},\mathbf{v}) = \sqrt[r]{\sum_{i=1}^{p}\left\|u_i - v_i\right\|^r}$
	Manhattan	$d(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{p}\left\|u_i - v_i\right\|$
	Chebychev	$d(\mathbf{u},\mathbf{v}) = max_{i=1}^{n}\left\|u_i - v_i\right\|$
	Camberra	$d(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{p}\frac{\left\|u_i - v_i\right\|}{\left\|u_i + v_i\right\|}$
Binary[60] Attributes $\{0,1\}$	Simple Matching Coefficient	$d(\mathbf{u},\mathbf{v}) = \frac{n_1 + n_2}{p}$
	Jaccard Similarity Coefficient	$d(\mathbf{u},\mathbf{v}) = \frac{n_1}{n_1 + n_3 + n_4}$
	Hamming Distance	$d(\mathbf{u},\mathbf{v}) = n_3 + n_4$
Nominal Scaled Attributes	Dissimilarity Coefficient	$d(\mathbf{u},\mathbf{v}) = \frac{p - o}{p}$
Ordinal Scaled Attributes	Similar to interval based attributes[61]	$z_i = \frac{u_i - 1}{U_k - 1}$

$\mathbf{u}$,$\mathbf{v}$: data vectors; p : number of attributes; r : positive integer; o : number of matches; n_1 : 1's in both vectors $\mathbf{u}$,$\mathbf{v}$; n_2 : 0's in both vectors; n_3 : 1's in $\mathbf{u}$ and 0's in $\mathbf{v}$; n_4 : 0's in $\mathbf{u}$ and 1's in $\mathbf{v}$, U_k : maximum value that attribute u_i can assume.

Table 2.13 - Distance between data instances (Andristos, 2002)

All the distance metrics should verify the following conditions:

[60] Choi *(Choi, 2010)* collected 76 binary similarity and distance measures.

[61] After the conversion to interval $[0,1]$ dissimilarity applied to interval scaled attributes must be computed.

$$
\begin{aligned}
&d(\mathbf{u},\mathbf{v}) \geq 0 \\
&d(\mathbf{u},\mathbf{v}) = 0 \ \textit{if and only if} \ \mathbf{u} = \mathbf{v} \\
&d(\mathbf{u},\mathbf{v}) = d(\mathbf{v},\mathbf{u}) \\
&d(\mathbf{u},\mathbf{z}) \leq d(\mathbf{u},\mathbf{v}) + d(\mathbf{v},\mathbf{z})
\end{aligned} \tag{2.71}
$$

$$\mathbf{u} = [u_1 ... u_p]^T;\ \mathbf{v} = [v_1 ... v_p]^T; \mathbf{z} = [z_1 ... z_p]^T$$

The vector of attributes may be composed of attributes with different natures (*mixed attributes* situation). According to Andristos *(Andristos, 2002)* in this case the distance between two data vectors $\mathbf{u}, \mathbf{v}$ can be given through the equation (2.72):

$$d(\mathbf{u},\mathbf{v}) = \frac{\sum_{j=1}^{p} \delta_j d_j}{\sum_{j=1}^{p} \delta_j} \tag{2.72}$$

the parameter $\delta_j = 0$ when u_j or v_j is missing, otherwise $\delta_j = 1$. The value of d_j depends on its type according to:

- Attribute u_j, v_j is binary or nominal: $d_j = 0$, if $u_j = v_j$, otherwise $d_j = 1$;
- Attribute u_j, v_j is interval scaled: $d_j = \left|u_j - v_j\right| / (max_j - min_j)$, where max_j is the maximum value of attribute j in the dataset and min_j is the minimum value of attribute j in the dataset;
- Attribute u_j, v_j is ordinal: $z_i = (u_i - 1) / (U_k - 1)$ must be calculated for u_j, v_j then the formula for interval scaled attributes $d_j = \left|u_j - v_j\right| / (max_j - min_j)$ must be applied.

Brief Survey of Clustering Algorithms

Several clustering algorithms have been developed to solve the unsupervised learning problem. However, these techniques can be divided in three main groups: *i*) partitional algorithms; *ii*) hierarchical algorithms; *iii*) density based algorithms *(Han, 2011)*.

Han defines partitional clustering as a class of algorithms *that construct k partitions of the data, where each cluster optimizes a clustering criterion, such as the minimization for the sum of squared distance from the mean within each cluster (Han, 2011)*.

K-means is a classic algorithm that belongs to the category of partitional clustering. Hammouda *(Hammouda, 2000)*, details the description of the algorithm considering that a set of N vectors $\mathbf{x}_i$, $i=1,...,N$ may be partitioned into K groups G_j, $j=1,...,K$ that contain the respective cluster centers $\mathbf{c}_j$, $j=1,...,K$ such that:

$$Q_j = \sum_{r:\mathbf{x}_r \in G_j} \left\| \mathbf{x}_r - \mathbf{c}_j \right\|^2 \tag{2.73}$$

where Q_j is the objective function to be minimized, i.e. the distance between the instances $\mathbf{x}_r \in G_j$ and the respective cluster center $\mathbf{c}_j$.

The partitioned groups G_j are defined by a $K \times N$ binary membership matrix, K clusters, N instances, where the respective elements $u_{ji}=1$ if the instance $\mathbf{x}_i$ belongs to group G_j, as presented in (2.74)[62]:

$$u_{ji} = \begin{cases} 1 \; if \; \left\| \mathbf{x}_i - \mathbf{c}_j \right\|^2 \le \left\| \mathbf{x}_i - \mathbf{c}_k \right\|^2, \; for \; k \ne j \\ 0 \; , \; otherwise \end{cases} \tag{2.74}$$

The centers of the different clusters can be updated as follows:

$$\mathbf{c}_j = \frac{1}{\left| G_j \right|} \sum_{r:\mathbf{x}_r \in G_j} \mathbf{x}_r \tag{2.75}$$

The algorithm may be described such as:

1. Choose K cluster centers to coincide with K randomly-chosen patterns or K randomly defined points inside the hypervolume containing the pattern set.
2. Assign each pattern to the closest cluster center.
3. Recalculate the cluster centers using the current cluster memberships.
4. If a convergence criterion is not met (e.g. reassignment of patterns to a new cluster, there is a decrease in squared error) go back to step 2 and repeat the process.

Figure 2.28 - k-means clustering algorithm (Jain, 1999).

Thus, the equations (2.73) and (2.74) implement step 2 of the algorithm described in Figure 2.28, step 3 is implemented through (2.75). K-means is

[62] The Euclidean distance can be replaced by other distance measure (Table 2.13).

extensively used in cluster analysis, although it presents several drawbacks: *i*) the difficulty in choosing the proper number of clusters centers; *ii*) the accuracy of the algorithm depends directly on the initialization of the clusters centers that must be a priori defined; *iii*) the algorithm is sensitive to outliers *(Mocian, 2009)*. There are several extensions to original k-means in order to minimize these flaws, e.g. k-means++, k-Medoids *(Witten, 2011) (Han, 2011)*. Fuzzy clustering and search techniques-based clustering algorithms are also important partitional clustering algorithms nonetheless they are not detailed in this thesis *(Xu, 2009)*.

Hierarchical algorithms create a multilevel hierarchy where clusters at one level are joined as clusters in the next level. The graphical representation of this hierarchical decomposition of the data instances is designated as dendrogram.

The methods used to decompose data instances hierarchically can be agglomerative (bottom-up) or divisive (top-down). The former assumes that each instance is a separate cluster which must be merged according to a specific distance measure. The stop condition is achieved when all the clusters belong to the same cluster. The divisive method works in reverse mode, it assumes that all data instances belong to the same cluster which must be split in disjoint clusters. The process iterates until each data instance belongs to a separate cluster or it reaches a stopping condition.

Jain *(Jain, 1999)* refers an additional classification of hierarchical algorithms based on the similarity measure between a pair of clusters: *i*) single link method, which defines the distance between two clusters as the minimum distance between all pairs of instances that belong to different clusters; *ii*) complete link method, that defines distance between two clusters as the maximum of all pairwise distances between instances in the two clusters. For both methods, the hierarchical agglomerative clustering algorithm can be described as:

1. Compute the proximity matrix containing the distance between each pair of clusters. Each pattern is perceived as a cluster.
2. Find the most similar pair of clusters (proximity matrix). Merge these clusters into one.
3. Update the proximity matrix to reflect this operation.
4. Stop if all instances belong to the same cluster. Otherwise, return to step 2.

Figure 2.29 - Hierarchical agglomerative clustering algorithm (Jain, 1999).

Subtractive clustering is a density based algorithm, since it groups data instances according to their density, i.e. the number of data instances in a specific

neighborhood. All the data instances are viewed as candidates for cluster centers, being assessed the respective density through the following formula:

$$\rho_i = \sum_{j=1}^{N} exp(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{(\chi / 2)^2}) \tag{2.76}$$

The parameter χ is a positive constant representing a neighborhood radius. A data point will have a high density value if it has many neighboring data points. The first cluster center is the one that presents the highest density value ρ_i. Then, the density values are updated as follows:

$$\rho_i = \rho_i - \rho_{\mathbf{c}_1} exp(-\frac{\left\|\mathbf{x}_i - \mathbf{c}_1\right\|^2}{(\chi_d / 2)^2}) \tag{2.77}$$

where the parameter $\rho_{\mathbf{c}_1}$ is the density value of the previous cluster center $\mathbf{c}_1$, χ_d is a positive constant that decreases the density in the neighborhood of $\mathbf{x}_i$. This density reduction is directly proportional to the proximity between the several data instances and the selected cluster center $\mathbf{c}_1$. The process iterates in order to find the next cluster center until an adequate number of clusters is identified. Data instances are assigned to the several clusters centers *(Hammouda, 2000).*

A clustering algorithm should verify some features: *i*) scalability, i.e. the ability to perform well with a large number of data instances; *ii*) ability to group mixed attributes; *iii*) handling of outliers; *iv*) insensitivity to the order of attributes in the data instances; *v*) ability to deal with a large number of attributes *(Andristos, 2002), (Han, 2011).* Clustering is a vast issue that cannot be fully explored in this thesis, however there are several comprehensive approaches that give a broad overview on clustering algorithms *(Han, 2011), (Witten, 2011), (Xu, 2009), (Fung, 2001).*

Subtractive clustering (density based algorithm) is the selected clustering algorithm to complement the dimensionality reduction procedure. These steps are critical to implement the methodology (grouping of patients) described in Section 3.6.

2.5 Validation

The validation phase is determinant to evaluate the potential clinical importance of the proposed methodology. Thus, this task must be as inclusive and accurate as possible. Although, it is important to refer that the definition of the validation

methodology was directly influenced by some limitations of the available real patient's datasets used in this work.

According to Steyerberg *(Steyerberg, 2009)*, validation can be organized in four main categories:

- *Apparent Validation*, when the training dataset is the same as the testing dataset. This validation may originate a biased performance assessment since the model parameters were optimized for that sample;
- *Split-Sample Validation*, where the dataset is randomly divided in two groups, one dedicated to develop the model and the other to validate the model. This option has a serious drawback since a random separation in two groups may not assure the right conditions for validation. This is critical namely in samples with a low event rate;
- Cross-validation is similar to the previous technique but in this situation the original dataset is divided in more sub datasets (e.g. k subdatasets, where $k-1$ are used to develop the model and the other is used for its validation). In this example, the validation procedure is repeated k times, in order to assure that all the elements of the original dataset are used at least once to validate the model. The overall performance is estimated from the average of all individual validations;
- Bootstrapping validation. In this situation bootstrap samples are drawn, with replacement[63], from the original sample being of the same size as the original sample. This approach is based on the assumption, that the original sample represents the population from which it was drawn. So resamples from the original sample represent approximately the same as what would be obtained with many samples directly pulled from the population. Statistics must be derived through the global analysis of the extracted bootstrap samples.

Due to the low event rate of the available real patient's dataset[64] (imbalanced datasets) the bootstrapping validation is particularly important for the implemented validation procedure. Therefore, it is broadly adopted in this thesis to reinforce the consistency of the validation results.

---

[63] Sampling with replacement means that after randomly drawing an observation from the original sample, that sample is put back before drawing the next observation.

[64] Santa Cruz Hospital dataset, Lisbon/Portugal; Leiria-Pombal Hospital Centre dataset, Leiria/Portugal

2.5.1 Bootstrapping Validation

Bootstrapping validation can be very useful to improve the reliability of the validation results, namely when the testing datasets are limited. According to Kirkwood *(Kirkwood, 2003)*, bootstrapping is based on the statement that if repeated samples are taken from the original sample, simulating the way the data are sampled from population, then these samples can be used to derive standard errors and confidence intervals. Several authors *(Wehrens, 2000) (Johnson, 2001) (Davison, 2006)* describe how bootstrapping validation may improve the reliability of the estimates of standard error and confidence intervals.

The standard error of the sample mean[65] (se) measures how precisely the population mean is estimated by the sample mean and is given by:

$$se = \frac{\sigma}{\sqrt{N}} \approx \frac{s}{\sqrt{N}} \tag{2.78}$$

where σ is the population standard deviation which is usually unknown. For that reason, the sample standard deviation s is applied to estimate the standard error. The larger the sample size N, the smaller is the value of se. The accurate assessment of the standard error is critical to the correct definition of the confidence interval.

Bootstrap strategy permits a different approach to the standard error calculation of the estimator $\hat{\theta}$ [66] considering B random samples (Figure 2.30).

1. Derive a random sample (with replacement) with the same dimension N as the original sample;
2. Compute the value of $\hat{\theta}$ for this bootstrap sample;
3. Repeat steps 1 and 2 until the B values of $\hat{\theta}$ have been computed;
4. Compute the standard deviation:

$$se_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum\left(\hat{\theta} - \bar{\hat{\theta}}\right)^2}$$

where $\bar{\hat{\theta}}$ is the mean of the B simulated values of the estimator $\hat{\theta}$.

Figure 2.30 – Bootstrapping procedure (Rossi, 2010).

[65] Standard deviation of the sampling distribution (frequency distribution of the individual sample mean).

[66] Estimator $\hat{\theta}$ of an unknown parameter θ, e.g. mean value.

Assuming that the sample mean $\overline{x}$ follows a normal distribution, the 95% confidence interval[67] is given by:

$$95\%CI = (\overline{x} - 1.96se,\ \overline{x} + 1.96se) \tag{2.79}$$

Bootstrap sampling allows the calculation of confidence intervals without the assumption that the statistic follows the normal distribution (or a different known distribution).

Considering that $\overline{X}$ is a random variable, μ is the actual value of the population mean and the confidence level is 95%, the values of o_1 and o_2 must verify the following expression:

$$P(\overline{X} \leq \mu + o_2) = 0.975 \quad \wedge \quad P(\overline{X} \leq \mu + o_1) = 0.025 \tag{2.80}$$

which is equivalent to:

$$P(\overline{X} - o_2 < \mu < \overline{X} - o_1) = 0.95 \tag{2.81}$$

the $\mu + o_2$ is approximately the 97.5$^{\text{th}}$ percentile of the distribution of $\overline{X}$ and $\mu + o_1$ is the 2.5$^{\text{th}}$ percentile of the distribution of $\overline{X}$.

Then, bootstrapping allows the direct estimates of o_1 and o_2 through the following procedure *(Johnson, 2001)*:

1. Extraction of the B bootstrapping samples.
2. For each sample i the statistic must be computed, e.g. mean value $\overline{x}_{(i)}$.
3. Estimation of the desired population percentiles. For instance, if $B = 1000$ are extracted, the estimates must be ordered $\overline{x}_{(1)} \leq \overline{x}_{(2)} \leq ... \leq \overline{x}_{(999)} \leq \overline{x}_{(1000)}$. The 2.5^{th} percentile is estimated through $\overline{x}_{(25)}$ and the 97.5^{th} percentile assumes the value of $\overline{x}_{(975)}$
4. The 2.5^{th} and 97.5^{th} percentiles must be respectively equal to $\mu + o_1$; $\mu + o_2$. The values of o_1 and o_2 are obtained, the value of μ is estimated by the original sample mean value $(\overline{x})$. The desired confidence interval is given by $(\overline{x} - o_2, \overline{x} - o_1)$.

Figure 2.31 - Derivation of 95% confidence interval.

[67] A confidence interval is an interval estimator that is designed to produce an interval of estimates that captures the unknown value of the parameter being estimated with a given probability. This probability of capturing the true value of the parameter being estimated is called the confidence level which is given by $(1-\alpha)\times 100\%$ *(Rossi, 2010)*.

Therefore it is possible to state that a $100\times(1-\alpha)\%$ bootstrap confidence interval for the parameter θ using the estimator $\hat{\theta}$ is given by:

$$(2\overline{\hat{\theta}}-\hat{\theta}_{upper}, 2\overline{\hat{\theta}}-\hat{\theta}_{lower}) \tag{2.82}$$

where $\hat{\theta}_{upper}$ is the $B(1-\alpha/2)$ order statistic of the bootstrap estimates and $\hat{\theta}_{lower}$ is the $B(\alpha/2)$ order statistic of the bootstrap estimate.

The following procedure (Figure 2.32) allows the standard error's estimation of the difference between the values of a given estimator for two different populations. The confidence intervals of the difference can be calculated as explained above in equation (2.82).

1. Derive B random samples (with replacement) with the same dimension as the original samples.
2. Compute the value of $\hat{\theta}_a, \hat{\theta}_b$ for these bootstrap samples.
3. Repeat steps 1 and 2 until the B values of $\hat{\theta}_a, \hat{\theta}_b$ have been computed.
4. Compute the standard deviation of the B simulated values of the estimators $\hat{\theta}_a, \hat{\theta}_b$ based on:

$$se_{\hat{\theta}_a-\hat{\theta}_b} = \sqrt{\frac{1}{B-1}\sum\left[(\hat{\theta}_a-\hat{\theta}_b)-(\overline{\hat{\theta}}_a-\overline{\hat{\theta}}_b)\right]^2}$$

where $\overline{\hat{\theta}}_a, \overline{\hat{\theta}}_b$ are the means of the B simulated values of the estimators $\hat{\theta}_a, \hat{\theta}_b$.

Figure 2.32 – Bootstrap procedure – comparison of two populations (Rossi, 2010).

The minimum number of bootstrap samples that achieve stable results is not consensual, however a significant number of authors consider that ideally 1000 bootstrap samples should be extracted *(Johnson, 2001) (Kirkwood, 2003) (Rossi, 2010)*. A deeper discussion on the bootstrapping issue can be found in *(Davison, 1997) (Wehrens, 2000)*.

2.5.2 Performance Assessment

The performance of a classifier is usually evaluated taking into account the confusion matrix presented in Figure 2.33:

	Positive (actual)	Negative (actual)
Positive (predicted)	*TP*	*FP*
Negative (predicted)	*FN*	*TN*

(**TP**) True Positive: patients with a positive test who were correctly diagnosed; (**FP**) False Positive: patients with a positive test who were incorrectly diagnosed; (**FN**) False Negative: patients with a negative test who were incorrectly diagnosed; (**TN**) True Negative: patients with a negative test who were correctly diagnosed.

Figure 2.33 - Confusion matrix.

Table 2.14 details some common metrics used to assess the performance of classifiers:

Parameter	Formula	Comments
Sensitivity (recall)	$\frac{TP}{TP + FN}$	Percentage of positive labeled instances (actual condition) that were predicted as positive.
Specificity	$\frac{TN}{TN + FP}$	Percentage of negative labeled instances (actual condition) that were predicted as negative.
Positive Predictive Value (Precision)	$\frac{TP}{TP + FP}$	Percentage of correct positive predictions.
Negative Predictive Value	$\frac{TN}{FN + TN}$	Percentage of correct negative predictions.
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Percentage of correct predictions.

Table 2.14 - Classifiers performance assessment.

When the datasets are imbalanced (frequent or rare outcome/events) accuracy is not a sensitive indicator of the models performance. This is a very common problem in real clinical data. In order to circumvent this problem Kubat *(Kubat, 1998)* proposed the geometric mean G_{mean} which considers the percentage of true cases correctly identified (SE - sensitivity) and the percentage of negative cases also correctly identified (SP – specificity) according to the following expression:

$$G_{mean} = \sqrt{SE \times SP} \tag{2.83}$$

Additionally $F_{measure}$ can also be used to measure the performance of the classifier:

$$F_{measure} = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.84}$$

The influence of the cut-off values in the performance of a binary classifier can also be assessed through a Receiver Operating Characteristic (ROC curve).

ROC curve is a plot of SE against $(1-SP)$ for different choices of the cut-off value *(Kirkwood, 2003)*. If the classifier is perfect (maximum discrimination capability) the $SE = 100\%$ and $SP = 100\%$, which means that the area under the curve (AUC) would be $AUC = 1$. On the contrary a classifier with minimum discrimination ability (unacceptable) only achieves $SE = 100\%$ with $SP = 0\%$ and conversely a $SP = 100\%$ implies a $SE = 0\%$, with an $AUC = 0.5$.

2.5.3 Hypothesis Tests

Statistical significance tests are very important when comparing the behavior of different classifiers operating with the same testing data set. They evaluate the evidence against the null hypothesis that is usually formalized as:

- H_0: Null hypothesis states that any difference of a given variable extracted from two datasets is due to chance or sample error;
- H_1: Alternative hypothesis, states that there is a reliable difference between the derived variables.

The significance level of a hypothesis test (*p-value*) is the probability of getting a difference at least as large as the one in the current sample if the classifiers have the same behavior. As stated by Kirkwood *(Kirkwood, 2003):*

- *The smaller the p-value the stronger is the evidence against the null hypothesis.*

Figure 2.34 presents an interpretation of the significance values, *p-values* lower than 0.05 are often reported as statistically significant to reject the null hypothesis.

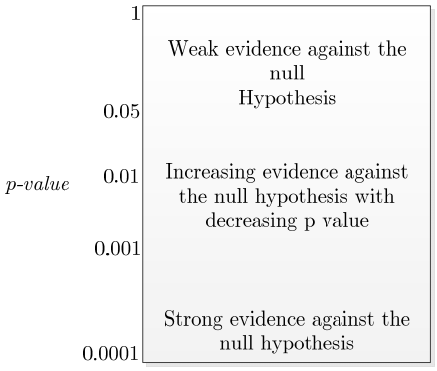


Figure 2.34 - Interpretation of p-value (Kirkwood, 2003).

The interpretation of the *p-value* may be complemented with the confidence interval (CI) of the variable under analysis. If the 95% confidence interval does not contain the null value, then the *p-value* must be smaller than 0.05. The opposite is also true, if the 95% CI includes the null value then the *p-value* must be greater than 0.05.

A hypothesis test never proves that the null hypothesis is true or false, since it only gives an indication of the strength of the evidence against the null hypothesis. Thus, two types of error can occur: *i*) type I error; *ii*) type II error.

Significance test	Null hypothesis is true (*actual*)	Null hypothesis is false (*actual*)
Reject null hypothesis	*Type I error*	*Correct decision*
Do not reject null hypothesis	*Correct decision*	*Type II error*

Table 2.15 - Types of errors (Kirkwood, 2003).

Type I and type II errors are closely related with the concepts of sensitivity and specificity. Type II error indicates a lack of significant difference between groups when in fact that difference exists (false negative). Therefore a test with high sensitivity has a low type II error rate. Type I error indicates that there is a significant difference between groups when in fact that difference does not exist (false positive). Thus, a test with high specificity has a low type I error rate.

Despite the occurrence of these errors (type I and type II), statistical significance tests are undoubtedly a very useful tool to extract more reliable conclusions on the comparison of the performances of different classifiers (groups of data). There are

several tests that must be adopted as they have different goals. The selection of the test also depends on several aspects such as the knowledge about the dataset, the number of dependent/independent variables, etc.

Type	Goal	P/Non-P	Name
One sample test	Compare one group to a hypothetical value	Non-parametric	Binomial Test
			Chi-square goodness-of-fit test
			Wilcoxon T
		Parametric	One-sample t test
Significance of group differences	Compare two unpaired groups	Non-parametric	Chi-square test of independence
			Fisher's exact test
			Wilcoxon / Mann Whitney U[68]
		Parametric	Unpaired samples t test
	Compare three or more unpaired groups	Non-parametric	Chi-square test of independence
			Kruskal-Wallis test
		Parametric	One-way ANOVA
			Factorial ANOVA
	Compare two paired groups	Non-parametric	McNemar test
			Sign test
			Wilcoxon signed-rank test
		Parametric	Paired samples t test
	Compare three or more paired groups	Non-parametric	Cochran's Q test
			Friedman two-way analysis of variance
		Parametric	One-way repeated-measures ANOVA
			Factorial repeated-measures ANOVA
Degree of relationship	Quantify association between two variables	Non-parametric	Kendall correlation
			Spearman correlation
		Parametric	Pearson correlation
			Partial correlation

Table 2.16 - Types of hypothesis tests[69] *(Sheskin, 2004).*

Table 2.16 shows part of a statistical test overview that systematizes the selection of the most suitable test for a given situation.

Levene's test is another statistical test that must be highlighted as it assesses the equality of variances in different samples. The null hypothesis assumes the homogeneity/equality of variances (homoscedacity). If there is strong evidence against the homogeneity of variances ($p\text{-}value<0.05$) it means that the differences

[68] These tests are equivalent. The test was first introduced by Wilcoxon, and afterwards by Mann and Whitney.

[69] Complete table accessed in June 2011: http://www.wiwi.uni-muenster.de/ioeb/en/organisation/pfaff/

cannot be explained by random factors therefore the null hypothesis must be rejected *(Sheskin, 2004)*.

Some of the metrics presented in Table 2.14 (sensitivity, specificity, accuracy and geometric mean) are thoroughly applied in the validation procedure. Furthermore, some tests that compare the mean/median between two unpaired groups (parametric Student's t-test/non parametric Mann Whitney U test) have particular importance. In some specific validation procedures a test to compare the means among more than two unpaired groups (one way ANOVA) is also applied. The Levene's test is also implemented to assess the equality of variances.

2.6. Conclusions

The main theoretical issues related with this thesis were addressed in this chapter, namely: *i*) current risk assessment tools for CAD and HF patients; *ii*) supervised classifiers and their application to modelling CVD risk assessment tools; *iii*) models' combination methodologies; *iv*) optimization methodologies (genetic algorithms); *v*) techniques to deal with missing risk factors; *vi*) dimensionality reduction and clustering techniques; *vii*) validation.

It is important to emphasize that different options could have been selected to implement the proposed methodologies. However, taking into account the specific conditions to develop this work (available datasets for validation, patients' condition (CAD/HF), etc.) as well as its main objectives (combine available information, ability to deal with missing risk factors, incorporation of clinical knowledge, etc.), the following aspects are particularly relevant in this thesis:

- This work addresses the combination of individual risk assessment tools for patients with cardiovascular disease (coronary artery disease, heart failure)[70]. Therefore, the identified models for secondary prevention (Table 2.2, Table 2.3) are particularly important for the validation procedure;
- Cox regression was explored as it was applied to the derivation of simulated models which is detailed in Section 3.4.1;
- The first step of the proposed methodology (Figure 1.2) is the common representation of individual risk assessment tools. Bayesian classifiers were

[70] The application of the proposed methodology to primary prevention is similar to secondary prevention.

selected to achieve this goal as they are suitable to accomplish the requirements of this thesis;

- Combination of individual models is the second step of the methodology presented in Figure 1.2. The implementation of the combination was performed through the direct combination of the parameters from the individual models (parameter/data fusion). This is a less explored approach to models' combination, although it seemed suitable for the required flexibility, e.g. incorporation of additional clinical knowledge, parameters adjustment. Afterwards, genetic algorithms were applied in this phase to adjust the parameters of the derived global model and consequently improve its prediction capability;
- An additional approach to enhance the performance of the risk assessment when compared to the one achieved by the current risk assessment tools is also proposed in this work. This methodology was based on the creation of patient groups. Dimensionality reduction techniques as well as unsupervised learning methods namely clustering algorithms are particularly important for its implementation;
- Validation is a critical phase of this work. This phase was performed as thoroughly as possible regarding the restrictions imposed by the available datasets. In this context, Bootstrapping validation was applied to reinforce the obtained results.

  The performance of the developed models was assessed through some of metrics detailed in Table 2.14 and complemented with statistical significance tests such as: *i*) Student's t-test for comparison of means between two unpaired groups, *ii*) Levene's test to compare variances between two unpaired groups, *iii*) One-way ANOVA to compare means between more than two unpaired groups.

  In some specific test cases additional non-parametric tests were performed (Mann Whitney U) to strengthen the results obtained from the parametric tests[71].

The next chapter gives a global perspective of the proposed methodology with a detailed explanation of each one of these aspects.

[71] This procedure was justified due to the reduced sample size.

3. Methodology

3.1 Introduction

The main goal of this thesis is to improve the CVD risk assessment based on currently available knowledge (current risk assessment tools). This enhancement is directly related with the increase of the risk prediction performance (SE/SP)[72] as well as with a set of characteristics that make this prediction more consistent.

In this context, this thesis aims to avoid/minimize some of the identified weaknesses of the current CVD risk assessment tools, namely: *i*) to consider the available knowledge provided by the current risk assessment tools; *ii*) to allow the consideration of a higher number of risk factors; *iii*) to cope with missing risk factors; *iv*) to incorporate empirical/additional clinical knowledge; *v*) to assure the clinical interpretability of the model; *vi*) to avoid the need of choosing a specific model as a standard model to be applied in the daily clinical practice; *vii*) to improve the performance of the risk assessment (SE/SP) when compared with the one achieved by the current risk assessment tools.

In order to reach these targets two methodologies are proposed: *i*) combination of individual risk assessment tools; *ii*) personalization based on grouping of patients.

These methodologies are presented as alternative methodologies as they use the current risk assessment tools based on different perspectives. However, it is important to mention that a global framework that merges these two methodologies should be explored in the ongoing research. Together, these two different perspectives may contribute towards the main objective of improving the current risk assessment.

The first methodology, combination of individual risk assessment tools (Figure 3.1), is detailed in Sections 3.2, 3.3 and 3.4.

[72] SE: sensitivity; SP: specificity.

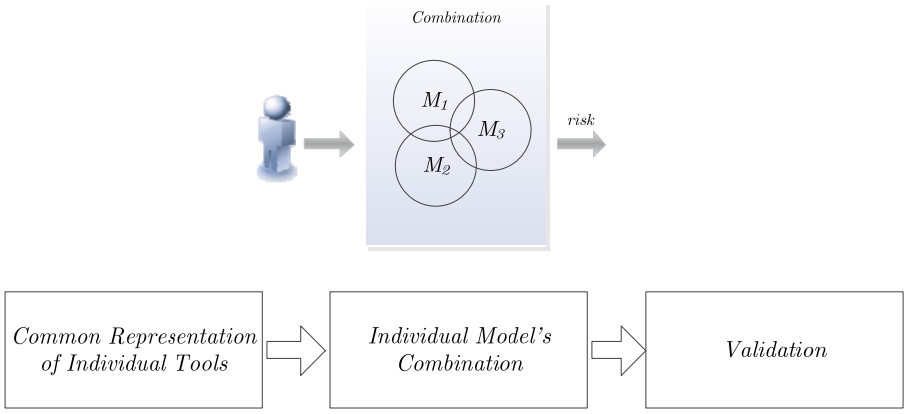


Figure 3.1 - Combination of individual risk assessment tools methodology.

Current risk assessment tools are diversely represented (charts, equations, etc.) which hinders their combination. Additionally, this diversity of representations is not suitable to achieve the goals of this work. The hypothesis proposed to solve this problem relied on the creation of a common representation of individual risk assessment tools that was based on a Bayesian classifier (naïve Bayes classifier). Therefore, Section 3.2 is dedicated to the first step of the methodology presented in Figure 3.1. The selection of the naïve Bayes classifier is justified along with the description of the derivation process of each classifier based on the respective current risk assessment tool.

The combination of individual models is the second step of the proposed methodology. The development of an efficient combination scheme is critical for the success and eventual clinical application of this work. In Section 3.3, two different combination schemes are explored. Both approaches can be included in the model parameter/data fusion category which was introduced in Section 2.2.2. The proposed combination may accomplish most of the mentioned goals. However it is not expected that this methodology alone will significantly improve the performance of the risk assessment achieved by the current risk assessment tools. Thus, an optimization procedure based on genetic algorithms is also depicted as it assumes great relevance for the adjustment of the global model's parameters (improvement of the model's performance). The strategy to deal with missing risk factors is also explored.

The last phase of the methodology (Figure 3.1) is validation that must be performed as comprehensively as possible. The available datasets for validation imposed several restrictions to this procedure. Hence two main scenarios for the application of the proposed combination methodology were created: *i*) theoretical

simulation applied to Heart Failure disease. Here, the derivation of simulated models was based on a real patients' dataset (the Trans-European Network Home Care Management System - TEN-HMS dataset *(Cleland, 2005)*) made available by the Castle Hill Hospital – Hull/UK ; *ii*) combination of current risk prediction tools for Coronary Artery Disease patients. This validation was based in two different datasets provided by the Leiria-Pombal Hospital Centre/Portugal and by the Santa Cruz Hospital – Lisbon/Portugal. Section 3.4 details these two validation scenarios. The metrics computed to assess the classifiers performance are identified as well as the statistical significance tests that were carried out to improve the reliability of the obtained results.

Closely related with these issues is the incorporation of clinical knowledge (Section 3.5) as it is a direct application of the combination of different risk assessment models represented as naïve Bayes classifiers.

Section 3.6 explores the personalization based on grouping of patient's methodology (Figure 3.2).

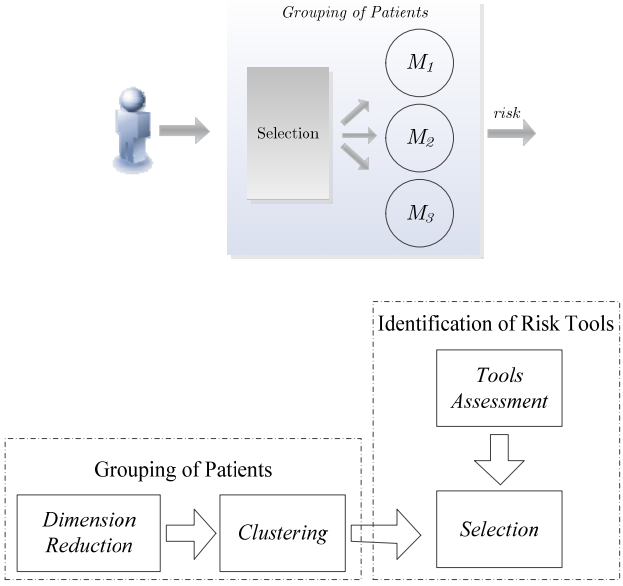


Figure 3.2 - Grouping of patients' methodology.

As mentioned, this methodology is based on the evidence that current risk assessment tools perform differently among different populations/groups of patients. Therefore, if the patients are properly grouped (clustered) it would be possible to find the best classifier for each group. The final classification of a given patient is

obtained through a selection process which considers the classification achieved by the individual risk assessment tool identified as the most suitable to classify the group of patients that the patient belongs to.

Section 3.7 addresses the validation of the personalization based on grouping of patients. Two validation scenarios are explored: *i*) simulation - theoretical individual models; *ii*) tools applied in clinical practice.

Section 3.8 systematizes the main concepts explored in this chapter.

3.2 Common Representation of Individual Tools

As previously referred, the diversity of representations of current individual risk assessment tools brings forth an additional difficulty to create a global model based on the combination of these individual elements. A hypothesis to solve this problem relies on the creation of a common representation that permits the direct combination of individual models. The selected classifier to implement this common representation was naïve Bayes classifier.

3.2.1 Naïve Bayes Structure

The selection of this classifier was based on: *i*) the particular features of the CVD risk assessment's problem; *ii*) the specific characteristics of the naïve Bayes classifier.

The CVD risk assessment intends to evaluate the risk of occurrence of an event (death, hospitalization, etc.) originated through cardiovascular disease within a specific period of time and given a set of risk factors. In this thesis, the risk is given through an output class where patients are classified according to two levels of risk (low/intermediate risk; high risk). The several risk factors (e.g. age, sex, hr[73], etc.) that should be statistically independent are the required inputs to compute that risk level.

As presented in Figure 3.3, the structure of the naïve Bayes classifier is well adapted to the specific characteristics of the problem under analysis. Moreover, the naïve Bayes classifier exhibits a set of characteristics that make it particularly

[73] heart rate.

suitable for the proposed methodology. These characteristics can be systematized as follows:

- Simple structure, which facilitates the creation of a proper combination scheme;
- Competitive performance with other classifiers (Table 2.5; Table 2.9). It is important to refer that some classifiers may present lower classification errors than naïve Bayes, e.g. Table 2.9. To circumvent this eventual lack of performance of the naïve Bayes classifier the proposed methodology comprises an optimization procedure;
- Ability to deal with missing risk factors. The inference mechanism of naïve Bayes (3.1) has an inherent ability to control the effect of missing risk factors.

$$P(C \mid \mathbf{X}) = P(C \mid X_1,...,X_p) = \alpha P(C)\prod_{i=1}^{p} P(X_i \mid C) \tag{3.1}$$

  The conditional probability table of a missing risk factor $P(X_i \mid C)$ is set to one which disables its influence in the final classification;
- Interpretability. The naïve Bayes parameters $P(X_i \mid C)$, $P(C)$ provide information about the probabilistic relationship between the several attributes X_i and the class of risk $c \in C$ as well as on the prior probability of the different risk classes. This probabilistic nature of parameters matches the reasoning required to establish a clinical diagnosis. Actually, it implies that the physician makes an inference which involves assessing the probability that a patient has a disease by the revealed symptoms *(Ayers, 2007)*;
- Computational efficiency, as presented in Table 2.10, this classifier has low complexity being the faster classifier not only in the training phase but also during the classification phase.

Therefore, the common representation of individual risk assessment tools was implemented based on naïve Bayes that presents the following structure:

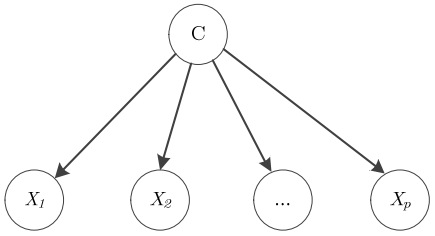


Figure 3.3 - Naïve Bayes structure.

As already introduced; this classifier is composed of only one parent (unobserved node: C) and several children (observed nodes: $X_1,...,X_p$). Its structure imposes a strong independence condition: all the attributes X_i are conditionally independent given the value of class C. In this work, as explained in Section 2.2.4, the eventual violation of the attribute's independence is limited. Although, the potential negative effect in the risk prediction originated by this violation is circumvented through an optimization procedure (genetic algorithm approach) that is carried out in the models' combination phase.

The final classification c is achieved based on the following equation:

$$c = \underset{c_j}{argmax}(\alpha P(c_j)\prod_{i=1}^{p} P(x_i \mid c_j)) \tag{3.2}$$

where c_j is a mutually exclusive class of C, x_i is the value of attribute X_i that belongs to the query instance $\mathbf{x}_q = [x_1,...,x_p]$. Thus, an instance $\mathbf{x}$ contains the values of a specific patient's attributes/risk factors (e.g. age, sex), c_j encodes a level of risk (e.g. low/high) and α is a normalization constant.

3.2.2 Naïve Bayes Parameters

The structure of naïve Bayes classifier is completely defined (Figure 3.3) as a result the learning process is restricted to parameters' learning. Thus, the model has to learn from the training data set, the conditional probability $P(X_i \mid C)$ of each attribute X_i given the class C as well as the prior probability $P(C)$ of the class C.

The process of representing an individual risk assessment tool as a naïve Bayes classifier can be systematized as follows:

- A training dataset (N instances $\mathbf{x} = [x_1 \dots x_p]$ composed of p attributes) is generated;
- This training dataset is applied to a given risk assessment tool in order to obtain a complete labelled dataset $J = \{(\mathbf{x}_1, c_1),...,(\mathbf{x}_N, c_N)\}$;
- Based on J and through the maximum likelihood estimation (Section 2.2.4.) it is possible to derive a naïve Bayes classifier that resembles the behavior of that specific risk assessment tool. The conditional probabilities can be calculated through the following expression:

$$P(X_i = x_i \mid C = c) = \frac{\sum_{1}^{N}(X_i = x_i \wedge C = c)}{\sum_{1}^{N}(C = c)} \tag{3.3}$$

The Laplace smoothing[74] is applied to avoid conditional probabilities with value 0. The prior probability $P(C)$ results directly from the distribution of the class values.

This process must be repeated to each one of the individual risk assessment tools that integrate the combination scheme.

3.2.3 Discretization

It is important to refer that this probabilities' estimation is reliable only when the attributes are categorical[75]. Hence the discretization of numeric attributes may have a great impact in the construction of the conditional probabilities tables and therefore in the performance of the classifier. The Equal Width Discretization (EWD) was the selected discretization method to allow the application of the maximum likelihood estimation to numeric attributes given by (3.3). However, in order to improve the clinical interpretability of the model, a discretization based on intervals with clinical significance was also tested in some attributes.[76]

3.3 Combination Methodology

The second step of the proposed methodology is the combination of individual models, i.e. the naïve Bayes classifiers that were created based on the risk assessment

---

[74] $\hat{\theta}_i = \frac{n_i + \alpha}{n + \alpha k}$; n_i: number of instances that assume the value i ; n: total number of instances; α: constant value (usually the value 1); k: number of the possible values of θ .

[75] Categorical variables comprise ordinal variables, nominal variables and dichotomous variables.

[76] Clinical guidelines define boundary values for some risk factors (e.g.: systolic blood pressure: less than 120 corresponds to normal values; [121–140] refers to the pre-hypertension category; more than 140 to hypertension stages (I, II). Accessed in June 2011: http://www.medicinenet.com

tools. Rather than to derive a completely new model, this combination methodology intends to create a classification system based on the incorporation of data from different sources (individual models)

As mentioned, combination methods can be grouped in two main categories: *i*) model output combination; *ii*) models parameter/data fusion. The proposed combination approach is included in the latter, since it takes advantage of the probabilistic reasoning as well as of the specific structure of the naïve Bayes classifier to implement the fusion of the individual models' parameters.

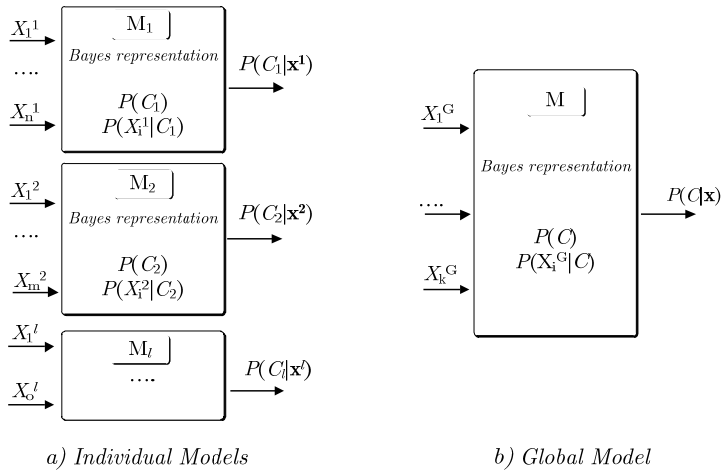


Figure 3.4 - Models' combination scheme.

Several individual classifiers $M_i \in \mathrm{M} = \{M_1,...,M_l\}$ are considered where each classifier is characterized by a specific conditional probability table $P(X_i^j \mid C_j)$, and by their respective prior probability of output class $P(C_j)$. $P(X_i^j \mid C_j)$ represents the conditional probability table of attribute i of model j, $P(C_j)$ the prior probability distribution of model j regarding a specific number of mutually exclusive classes, $\mathbf{x}^j \subseteq \mathbf{x} = [x_1 \ldots x_p]$ is the input instance considered by the model j (risk factors considered by j that are a subset of the p risk factors).

Regardless of the fusion approach, the model selection criterion to integrate the combination scheme highly influences the global classification performance. According to the implemented condition, the information of a given model is considered if there is at least one of its inputs available. Moreover, some risk factors may be considered by more than one model, while other inputs belong only to a single model. Therefore, the classification of the global model is dependent on the availability of input risk

factors as well as on the selection criteria to define the individual models that should be included in the combination scheme.

Additionally, to allow the combination of different individual models the following condition has to be verified:

- Individual models have the same number of output levels (e.g. "low/intermediate", "high").

This restriction that was applied in the proposed combination scheme ensures that models share the same risk assessment goal. The number of output levels is defined according to the requirements of the specific risk assessment.

In this thesis two risk levels were defined since the main clinical goal is the identification of the high risk patients. The clinical partner validated this approach where two risk classes were considered for classification purposes:

- *The reduction of output categories (low risk/high risk) is correct. In fact, the aim of cardiologist in clinical practice is to discriminate between high risk patients and low risk patients. In a clinical perspective, the identification of intermediate risk patients is not so significant.*

The combination strategy based on the identification of two risk classes can also be applied to a multiclass classification where the number of output risk classes is higher than two. Multiclass classification problems can be decomposed into multiple binary classification procedures whose outputs are combined to generate the final classification. This decomposition followed by the reconstruction phase may be implemented through two main techniques: *i*) coding matrix; *ii*) hierarchical algorithms. Several authors *(Allwein, 2000) (Schwenker, 2001)* detail these decomposition techniques nonetheless this topic is not explored as it is beyond the scope of this thesis.

Integrated in the category *models parameter/data fusion*, two different approaches were tested to perform the fusion of the individual statistical models: *i*) individual models parameters' union; *ii*) individual models parameters' weighted average.

3.3.1 Individual Models Parameters' Union

This models' fusion strategy considers that the global model is formed based on the union of several individual models.

Based on the rules of probability, prior probability and conditional probabilities of the global model can be derived as follows:

$$P(C) = P(C \cap (C_1 \cup C_2 \cup ... \cup C_l) = P\left(\bigcup_{i=1}^{l} (C \cap C_i)\right) \tag{3.4}$$

Where $P(C_i)$; $1 \le i \le l$ represent the risk distributions of the l individual models[77] and $P(C)$ the prior probability of risk of global model.

The equation (3.4) can be developed based on the rule of addition[78]:

$$\begin{aligned} P(C) = &\sum_{i=1}^{l} P(C \cap C_i) - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} P(C \cap C_i \cap C_j) + \sum_{i=1}^{l-2} \sum_{j=i+1}^{l-1} \sum_{k=j+1}^{l} P(C \cap C_i \cap C_j \cap C_k) \\ &+ ... + (-1)^{l+1} P(C \cap C_1 \cap C_2 \cap ... \cap C_l) \end{aligned} \tag{3.5}$$

which can be simplified:

$$\begin{aligned} P(C) = &\sum_{i=1}^{l} P(C_i) - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} P(C_i \cap C_j) + \sum_{i=1}^{l-2} \sum_{j=i+1}^{l-1} \sum_{k=j+1}^{l} P(C_i \cap C_j \cap C_k) \\ &+ ... + (-1)^{l+1} P(C_1 \cap C_2 \cap ... \cap C_l) \end{aligned} \tag{3.6}$$

assuming that individual models are statistically independent[79], it is possible to state:

$$\begin{aligned} P(C) \underset{C_i independent}{=} &\sum_{i=1}^{l} P(C_i) - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} P(C_i)P(C_j) + \\ &\sum_{i=1}^{l-2} \sum_{j=i+1}^{l-1} \sum_{k=j+1}^{l} P(C_i)P(C_j)P(C_k) + ... + (-1)^{l+1} P(C_1)P(C_2)...P(C_l) \end{aligned} \tag{3.7}$$

The conditional probabilities that form the conditional probability tables can be determined using a very similar reasoning:

$$\begin{aligned} P(X_i^j \mid C) &= \frac{P(X_i^j \cap C)}{P(C)} = \frac{P((X_i^j \cap C) \cap (C_1 \cup C_2 \cup ... \cup C_l)}{P(C)} \\ &= \frac{P((X_i^j \cap C \cap C_1) \cup (X_i^j \cap C \cap C_2) \cup ... \cup (X_i^j \cap C \cap C_l))}{P(C)} \\ &= \frac{P\left(\bigcup_{k=1}^{l} (X_i^j \cap C \cap C_k)\right)}{P(C)} \end{aligned} \tag{3.8}$$

---

[77] Individual models have the same number of mutually exclusive output classes.

[78] $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

[79] A set of events $A_1 ... A_n$ is statistically independent if: $P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$

where X_i^j is the attribute i of the individual model j and $P(X_i^j \mid C)$ is the conditional probability of X_i^j given C.

$$P(X_i^j \mid C) = \frac{1}{P(C)}[\sum_{k=1}^{l} P(X_i^j \cap C \cap C_k) - \sum_{k=1}^{l-1} \sum_{m=k+1}^{l} P(X_i^j \cap C \cap C_k \cap C_m) + ... \\ + ... + (-1)^{l+1} P(X_i^j \cap C \cap C_1 \cap C_2 \cap ... \cap C_l)] \tag{3.9}$$

as:

$$P(X_i^j \cap C \cap C_k) = P(X_i^j \mid C \cap C_k) P(C \cap C_k) = P(X_i^j \mid C_k) P(C_k)$$

$$P(X_i^j \cap C \cap C_k \cap C_m) = P(X_i^j \mid C \cap C_k \cap C_m) P(C \cap C_k \cap C_m) = P(X_i^j \mid C_k \cap C_m) P(C_k \cap C_m) \\ \underset{C_i \ independent}{=} P(X_i^j \mid C_k) P(X_i^j \mid C_m) P(C_k) P(C_m) \tag{3.10}$$

Then, $P(X_i^j \mid C)$ can be given by:

$$P(X_i^j \mid C) = \frac{1}{P(C)} \left[\sum_{k=1}^{l} P(X_i^j \mid C_k) P(C_k) - \sum_{k=1}^{l-1} \sum_{m=k+1}^{l} P(X_i^j \mid C_k) P(X_i^j \mid C_m) P(C_k) P(C_m) + ... + \right. \\ \left. (-1)^{l+1} P(X_i^j \mid C_1) P(X_i^j \mid C_2) ... P(X_i^j \mid C_l) \prod_{k=1}^{l} P(C_k) \right] \tag{3.11}$$

This approach has some important drawbacks that can originate biased results. Actually, the union of models implies that the global model's prior probability of risk $P(C)$ is always higher than the maximum probability of the individual models. The exception occurs when individual models' have the same output. The same reasoning can be taken to analyze the conditional probability tables $P(X_i \mid C)$.

Additionally, the union of the individual models does not consider the potential differences that can occur in the performances of the different classifiers towards a specific population. Thus, all the models have the same importance (weight) which may also increase the bias of the global classification. Initially this approach seemed a valid strategy to combine models, although based on some preliminary tests this first studied hypothesis for model's combination was discarded.

3.3.2 Individual Models Parameters' Weighted Average

This innovative combination methodology relies on the evidence that current cardiovascular risk assessment tools register different behaviors when applied to the classification of a specific population data set. In fact, the same tool may perform diversely in different testing datasets (some results that confirm this evidence are presented in Section 4.3.3.). Therefore, the combination strategy must be able to assign different weights for the individual models according to their performance in a specific dataset.

Moreover, the combination scheme must be prepared to accommodate different individual model selection criteria which have a direct influence in the set of individual models that integrate the combination scheme.

Some risk factors (model inputs) may be considered by more than one model, while other inputs belong only to a single model. The proposed method assures that a CPT calculation for a variable that is used by more than one model has to consider the information provided by those models. In contrast, if a variable only belongs to one model, the CPT table has to match the respective individual CPT.

The individual models parameter's weighted combination of the individual models is performed based on the following expressions:

$$\begin{aligned} P(C) &= \sum_{j=1}^{l} P(C_j) \times \frac{w_j}{\Gamma} \quad where \quad \Gamma = \sum_{j=1}^{l} w_j \\ P(X_i \mid C) &= \sum_{j=1}^{b} P(X_i^j \mid C_j) \times \frac{w_j}{\vartheta} \quad where \quad \vartheta = \sum_{j=1}^{b} w_j \end{aligned} \tag{3.12}$$

Where l is the number of individual models, b is the number of individual models that contain the attribute $X_i \in \{X_1,...,X_p\}$, C_j denotes each individual model, w_j is the weight of model j.

This combination scheme is the basis of the combination methodology proposed in this work. Actually, the combination defined in (3.12) is flexible which permits to implement a combination strategy that depends on the characteristics of each specific combination, namely it:

- Permits to assign to each individual model a different weight. The weight assigned to each model should be set according to the respective performance. The weights' definition may be done iteratively based on a set of test cases;
- Allows disabling a specific model. In this way, different individual model selection criteria to integrate the combination scheme may be implemented.

3.3.3 Optimization

As referred, naïve Bayes classifier often presents higher classification errors than other classifiers (e.g. semi naïve methods). This eventual lack of performance is addressed by the proposed combination scheme which includes an optimization procedure. It intends to adjust the models' parameters that result from the combination strategy in order to improve the performance of the global model.

The application of genetic algorithms seemed appropriate as they can be applied to both constrained and unconstrained optimization problems where the objective function $f(\mathrm{x})$ is nondifferentiable or highly nonlinear *(Eiben, 2003)*.

Here, the application of genetic algorithm (GA), focuses on $P(X_i \mid C), P(C)$ (probabilities) that are the parameters of the global model originated through the individual models parameters' weighted average method (Section 3.3.2).

The optimization procedure cannot distort the information provided by the individual risk assessment tools which is the basis of the global model parameters $P(X_i \mid C), P(C)$ definition. In this context, the adjustment of the global model's parameters must be constrained to the neighbourhood of the initial values that were calculated through (3.12).

Considering the values of $P(X_i \mid C)$, this constraint is given by:

$$-\varphi \times P(X_i = x_i^k \mid C = c_j) \leq \delta_{kj} \leq \varphi \times P(X = x_i^k \mid C = c_j) \tag{3.13}$$

The parameters to be optimized δ_{kj} denote the allowed variation on the probability of the category k of attribute X_i given the output class j (risk level), φ is the value of the neighbourhood that is adjusted experimentally.

This restriction may reduce the efficiency of the optimization algorithm, although it assures that the optimization procedure does not ignore the knowledge provided by the original models, i.e. it assures the clinical significance of the model.

Therefore, considering three possible categories for the attribute X_1, $\{x_1^1, x_1^2, x_1^3\}$ and two mutually exclusive risk classes $\{c_1, c_2\}$ for the output C, the conditional probability table is defined by a 3×2 matrix, as shown in equation (3.14).

$$\begin{bmatrix} P(X_1 = x_1^1 \mid C = c_1) & P(X_1 = x_1^1 \mid C = c_2) \\ P(X_1 = x_1^2 \mid C = c_1) & P(X_1 = x_1^2 \mid C = c_2) \\ P(X_1 = x_1^3 \mid C = c_1) & P(X_1 = x_1^3 \mid C = c_2) \end{bmatrix} \tag{3.14}$$

Then the optimization procedure uses the previous information and is conducted in the neighbourhood of the initial values, as represented in the following expression:

$$\begin{bmatrix} P(X_1 = x_1^1 \mid C = c_1) \pm \delta_{11} & P(X_1 = x_1^1 \mid C = c_2) \pm \delta_{12} \\ P(X_1 = x_1^2 \mid C = c_1) \pm \delta_{21} & P(X_1 = x_1^2 \mid C = c_2) \pm \delta_{22} \\ P(X_1 = x_1^3 \mid C = c_1) \pm \delta_{31} & P(X_1 = x_1^3 \mid C = c_2) \pm \delta_{32} \end{bmatrix} \tag{3.15}$$

The first step of a GA application is to define how to represent the individuals for the required adjustment. In this case, the individuals are represented as real numbers codifying the variation of each parameter. The size of the population (number of individuals) is usually set to a higher value than the size of an individual (number of parameters to optimize).

As mentioned, the aim of this optimization procedure is to improve the performance of the risk prediction provided by the global model, i.e. maximize the sensitivity and specificity of the risk prediction. An evaluation step must be defined in order to assign a quality measure to each candidate solution considering the defined goal. Thus, the selected evaluation step is composed of two functions (f_1, f_2 *multiobjective optimization*[80]) since the optimization attempts to maximize simultaneously the specificity and the sensitivity of the global model.

The criteria f_1, f_2 were defined (3.16) in order to transform the maximization of specificity and sensitivity into a minimization problem as presented in (2.59).

[80] Multiobjective optimization is applied when a single objective with several constraints does not adequately represent the optimization problem. In multiobjective optimization there is a vector of objective functions $\mathrm{f} = [f_1 ... f_n]$, where a tradeoff between objectives must be found. In this context a noninferior solution (also designated *Pareto optima*) is one solution in which an improvement in one objective implies a degradation of another objective.

$$f_1 = 1 - \frac{TP}{TP + FN}; \quad f_2 = 1 - \frac{TN}{TN + FP} \tag{3.16}$$

TP : True Positive; TN : True negative; FN : False negative; FP : False Positive

The parent's selection function is implemented through the roulette wheel algorithm (Figure 2.19).

Variation operators are also very important for the operation of genetic algorithms. In this optimization, the selected crossover operator is the uniform crossover (Figure 2.22). A Gaussian mutation operator is applied to assure the required mutation.

Survival selection is implemented based on the fitness of individuals of current generation as well as on the fitness of individuals of the respective offspring (fitness based replacement).

Finally, several termination conditions were tested. The condition that verifies if the fitness improvements remain under a threshold value during a given period of time is adopted in this optimization.

It is important to emphasize, that the tuning[81] of a genetic algorithm can be challenging. In this work the tuning was performed based on an extensive set of experiments[82].

3.3.4 Missing Information

Missing information (risk factors) is a very frequent problem in health records *(Khanna, 2005)*. The proposed combination strategy addresses this evidence, through the particular characteristics of the inference mechanism of the naïve Bayes classifier (3.1).

This option is possible since the classifier that results from the combination of individual naïve Bayes classifiers is also a naïve Bayes classifier. Thus, when there is a missing risk factor X_i. the values of the respective conditional probability table $P(X_i \mid C)$ are replaced by value 1.

[81] Definition of the genetic algorithm parameters such as: *population size*; *population initial range*; *number of generations*; *crossover function; mutation function; crossover rate; mutation rate*; *stopping condition.*

[82] Genetic algorithms were implemented based on the Global Optimization Toolbox, Matlab.

This replacement has limited effects in the prediction performed by the classifier which allows a lower classification error than some of the common techniques used to deal with missing information. The inherent capability to deal with missing information is one of the advantages of Bayesian classifiers as detailed in Table 2.5.

3.4 Validation of the Combination Methodology

The potential clinical relevance of the proposed methodology depends directly on the validation procedure reliability. This step was performed so comprehensively and detailed as possible, which originated two different validation scenarios:

- The first validation scenario (Simulation - Theoretical Individual Models) was created due to limitations of the available dataset. In fact, the limitation of the TEN-HMS dataset[83] did not enable the direct use of current risk assessment tools (Table 2.2).

  In order to circumvent this additional difficulty, simulated models were derived and afterwards combined. The simulation of individual models did not involve any modification in the proposed combination scheme. Finally, it must be stressed that the creation of the individual theoretical models was guided by a real patients' dataset;
- Actual validation, where the proposed combination methodology was implemented with tools that are currently applied in regular clinical practice. This combination was tested with real patients' datasets[84].

3.4.1 Simulation – Theoretical Individual Models

This validation scenario was severely influenced by some limitations of the TEN-HMS dataset, as it does not have enough variables that allow the combination of current risk assessment tools specific for Heart Failure patients.

---

[83] TEN-HMS dataset (426 patients) made available by Castle Hill hospital, Hull UK. The complete description of this dataset is given in Section 4.1.1.

[84] Datasets made available by Santa Cruz Hospital, Lisbon (460 patients) and by Leiria-Pombal Hospital Centre, Leiria (99 patients). These datasets are detailed in 4.2.2.

In this context, the first step of the proposed methodology (Figure 3.1) had to be adapted in order to accommodate the derivation of the individual models required to support the proposed combination strategy (Figure 3.5).

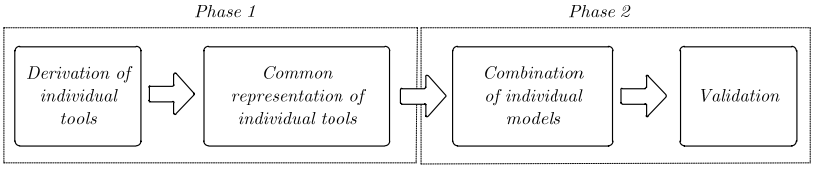


Figure 3.5 - Proposed methodology (simulation)

Cox regression[85] has been selected to derive the individual models. However the generation of each one of the individual models has not been a trivial issue, since the following questions must be solved:

- Which variables/risk factors must be considered?
- How to group those variables, i.e. which specific individual models should be created? How many models should be created?

The selection of the variables to incorporate a Cox model must be performed considering that variables have to be significantly related to survival time and they should not correlate strongly with each other. Two options are commonly addressed for the selection of variables:

- Based on a statistical analysis: the goal is to identify the most significant risk factors among the variables identified in the literature as being important predictors of outcome in patients with a specific disease. That selection might be optimized afterwards through some techniques such as Recursive Feature Elimination *(Guyon, 2002)*;
- Based on a set of variables used by a specific risk tool. It is assumed that the selection of significant variables has already been identified and validated.

The second option is followed in this work. Given the available variables in the TEN-HMS dataset, and considering some of the most significant risk assessment tools for death prediction in Heart Failure patients (Table 2.2), particularly the Senni model *(Senni, 2006)*, twelve variables were identified in this approach. These attributes (risk factors) $\{X_1,...,X_p\}$ formed the global variable space required to derive the simulated models. The derivation of individual models (simulated models) creates an additional challenge related with the identification of the number of

[85] The most commonly used model to analyse survival data is the Cox proportional hazards model. Accessed in January 2011: http://www.healthknowledge.org.uk

models to be combined as well as their respective inputs (risk factors). Kutner *(Kutner, 2004)* supports the idea that:

- If the mean of the estimate error of all samples of each model are similar the validation information does not invalidate the model application to other samples belonging to the same study universe.

This is the basis of the proposed algorithm in order to choose the different subsets of the global variables space $\{X_1,...,X_p\}$, from which the individual models were derived through Cox Regression (Figure 3.6).

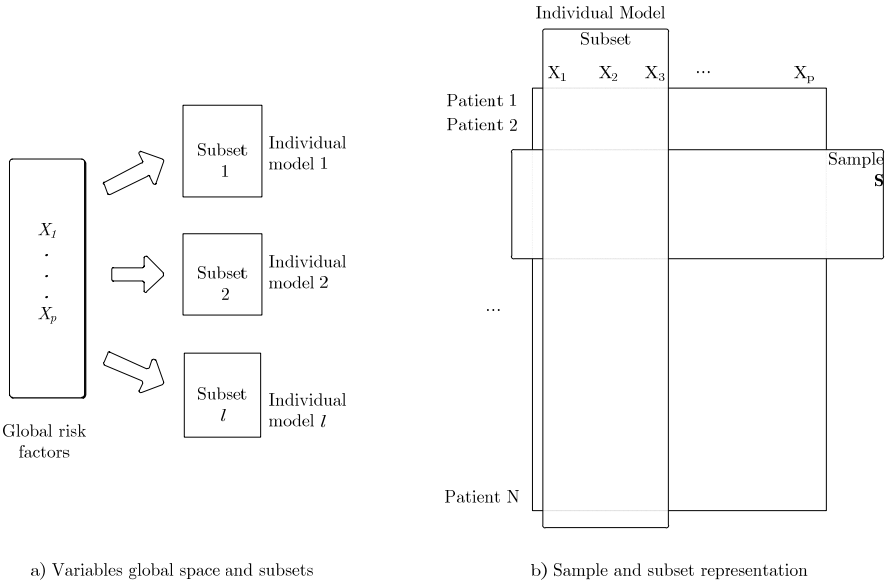


Figure 3.6 - Individual models' derivation.

Several subsets from the initial space of variables $\{X_1,...,X_p\}$ were created, Figure 3.6, each one corresponding to a distinct individual model M_i comprising of a subset of variables $S_{M_i} \subseteq \{X_1,...,X_p\}$. The number of variables of each model (subset) was defined between predefined limits. This procedure originated a total number of l candidate models to integrate the combination scheme.

As suggested by Kutner *(Kutner, 2004)*, instead of deriving each model M_i using a unique dataset, different parameterizations for each model M_i should be obtained using different subsets of the patients dataset. Then, from these different parameterizations, the most appropriate M_i models are selected. In this work, a number of distinct samples of patients $(S = 10)$ were created based on the available

dataset, Figure 3.6 b), in order to derive the distinct individual Cox regression models: $M_{i,s}$; $i = 1,...,l$; s=1,...,10. Each sample has the same length, i.e., the same number of patients and each patient may belong to more than one sample. Therefore, for each model M_i, $S = 10$ different Cox regression models were created, resulting in a total of $l \times S$ different models. However, it should be noted that only the individual models that present the lowest accuracy's variance were selected to the combination phase.

The selection of individual Cox models $M_{i,s}$ was based on the comparison with the performance of a complete Cox model. This complete Cox model has been obtained considering all risk factors and all patients of the available dataset and it has also been derived through Cox Regression technique.

In particular the mean and standard deviation of errors between individual models and the complete model were analyzed. Basically, models $M_{i,s}$ $s = 1,..,10$ that presented the lowest accuracy's variance among the respective M_i models were selected as having high potential for the combination phase.

The number of selected models was defined regarding the minimum number of models required to perform a number of combinations that assured the statistical significance of the obtained results

According to Figure 3.5, after the conclusion of this first step, the common representation of individual models based on naïve Bayes classifier has been completed and the combination strategy has been applied.

This simulation investigated the combination of only two individual models. This number is justified by clinical practice aspects. In effect, a physician has typically to deal with two or three distinct models. Moreover, the extension of the current approach to a higher number of models is straightforward.

Finally, it is important to emphasize that in this specific validation the real patient dataset was the basis to the derivation of individual models. Training and testing dataset were simulated based on the respective variables' values available in literature.

Performance Assessment

Specificity, sensitivity and accuracy (Table 2.14) were computed separately to the individual models (simulated models) as well as to the global model that resulted from the combination scheme. Additionally, the assessment of the global model's

performance was made before and after the optimization procedure based on genetic algorithms operation.

The *true data* needed to compute those metrics was obtained through the complete Cox model. A set of generated testing instances $\Upsilon = \{\mathbf{x}_1,...,\mathbf{x}_N\}$ was applied to that model in order to obtain the complete labelled testing dataset $D_T = \{(\mathbf{x}_1, c_1^T),...,(\mathbf{x}_N, c_N^T)\}$.

Some statistical significance tests were carried out to obtain more reliable conclusions about the performance of the models. The applied tests were: *i*) Student's t-test; *ii*) Levene's test; *iii*) Mann Whitney U test; *iv*) One-Way analysis of variance (ANOVA).

A Student's t-test (Table 2.16) was carried out to evaluate the statistical significance of the differences between the assessed metrics' mean values provided by the different classifiers. As referred, this test assumes a null hypothesis (H_0) that can be defined as:

- H_0: The mean values of two datasets are equal *vs.* H_1: The mean values of two datasets are not equal.

If there is strong evidence against the null hypothesis (*p-value*<0.05) it means that there is strong evidence (high probability) that the different classifiers have an effect (positive/negative) in the risk prediction. Based on this test, the global model after optimization was compared with: *i*) the global model before optimization; *ii*) the individual models.

Levene's test has been performed to evaluate the following null hypothesis:

- H_0: The variances of two datasets are equal (homogeneity of variances) *vs.* H_1: The variances of two datasets are not equal.

This test is critical to assess the homogeneity of variances, and identify those situations where it cannot be assumed.

The Mann Whitney U test, a non-parametric test, was also applied in order to confirm the results obtained with the parametric test. Actually, the selected parametric test (Student's t-test) can be safely applied when the data verifies the following assumptions: *i*) data is normally distributed; *ii*) statistical independence exists between groups of data; *iii*) there is homogeneity of variance between groups of data *(Dowdy, 2004)*. However, in this case there is a deviation from the normality assumption which can reduce the reliability of the test results. This violation may be

particularly significant as the sample size is somewhat reduced (approximately the minimum value that allows the application of the Central Limit Theorem)[86]. The Mann Whitney U test implements the following null hypothesis:

- H_0: The medians[87] between the two datasets are equal *vs.* H_1: The medians of two datasets are not equal.

One-Way analysis of variance (ANOVA) was also applied to reinforce the obtained results. According to Kirkwood *(Kirkwood, 2003), ANOVA is used to compare the mean of a numerical outcome variable in the groups defined by the exposure level with two or more categories. It is called one-way as the exposure groups are classified by just on variable.* This method determines how much of the overall variation in the outcome is attributed to the differences between the group means (more than two groups). The two main assumptions to apply the analysis of variance are: *i*) the outcome is normally distributed; *ii*) there is homogeneity of variances among the groups. According to several authors *(Kirkwood, 2003) (Rossi, 2010)* moderate departures from normality assumption may be safely ignored, while the unequal variances may be critical.

Missing Risk Factors

The final validation procedure performed with the combination of simulated individual models is related with the ability to deal with missing risk factors. As mentioned, the lack of input information is a very frequent problem in medical records *(Khanna, 2005)* or at the moment of assessing the risk, so this type of evaluation is very relevant and must be carried out.

Three different situations were compared: *i*) Bayesian global model that results from the combination scheme without any replacement of the missing variables; *ii*) Bayesian global model that results from the combination scheme with replacement of the missing variables; *iii*) global Cox model with replacement of the missing variables.

The continuous missing variables were replaced by the respective mean values, in the case of Boolean variables their value were successively replaced by 0 and 1 values.

---

[86] A large number of authors consider that the minimum sample size to apply the Central Limit Theorem is approximately $N > 30$, however depending on the author this number may vary between $25 < N < 40$.

[87] Median is the 50th numerical value that separates the higher half of a sample from the lower half.

In this case, the statistical validation was applied to the classifiers' accuracy and it was similar to the previously explained validation, i.e. based on: *i*) Student's t-test; *ii*) Levene's test; *iii*) Mann Whitney U test; *iv*) One-Way analysis of variance (ANOVA).

3.4.2 Tools Applied in Clinical Practice

This validation procedure focuses on the evaluation of the performance of the global model that is originated through the combination of tools that are applied in the daily clinical practice.

The first step is the selection of individual tools that are adequate to predict the risk of a CVD event regarding a specific disease. In this particular case, some current tools suitable to predict risk in coronary artery disease (CAD) patients have been selected. This selection process was supervised by the clinical partners that collaborated in this work.

The second step of the proposed methodology[88] (Figure 3.1) was applied similarly to the previous validation scenario.

The different real patient testing datasets were made available by two Portuguese hospitals. These testing datasets provided the *true data* required to compute all the metrics applied in the performance assessment. The training dataset required to generate the parameters to represent the individual Bayesian classifiers was derived based on proper values available in literature. This approach represents a significant difference in relation to the previous situation (simulation) where the real patient dataset is exclusively used to generate the individual models.

Performance Assessment – Individual Tools

The first stage was the assessment of the performance of the selected individual tools on the specific testing datasets. As referred, the performance of current tools may differ depending on the testing population, thus the information obtained in this phase provided the essential knowledge to adjust the weights of individual tools.

[88] The combination of the individual models was implemented as explained in Section 3.3.2.

Sensitivity, specificity, area under the ROC curve and geometric-mean were the metrics assessed. The latter was evaluated since both testing datasets are severely imbalanced due to their reduced event rate.

Performance Assessment – Global Model

The second step of this validation consisted in the assessment of the performance of the global model (after combination) when different weights were assigned to the individual models. This allowed the identification of the best weights combination to derive the global model.

For all testing datasets, the Bayesian global model has been compared with a voting model. This alternative combination approach was selected since it is the most likely to be applied by the physician in the daily clinical practice.

Likewise a global assessment was carried out, where the performance of the global Bayesian was compared not only with the voting scheme, but also with the individual models. This validation was important to determine if the combination strategy improved the final classification or otherwise was worse than any of the other models (individual/voting model). For all the testing datasets, this assessment was performed comparing the Bayesian global model with each one of the other models.

In order to increase the statistical significance of the obtained results, bootstrapping validation was employed which allowed the derivation of confidence intervals of the metrics assessed. Parametric statistical significance tests (Student's t-test, Levene's test) were also executed to increase the reliability of the conclusions extracted from this comparison. Analysis of variance (ANOVA) was introduced to provide a global perspective of the relationships among the several classifiers.

The evaluation of the optimization process was developed following the same approach. The performance of Bayesian global model before and after the optimization was compared. Statistical significance tests were applied to support the respective conclusions.

Missing Risk Factors

The ability of the Bayesian global model to deal with missing risk factors was also assessed. The effect of missing risk factors was evaluated, on all metrics that were computed considering the different testing datasets.

Therefore, each variable was successively removed. For each variable, the performance of three different models was evaluated: *i*) Bayesian global model before optimization; *ii*) Bayesian global model after optimization; *iii*) Voting[89].

The different models' performance was compared based on parametric statistical significance tests (Student's t-test, Levene's test) that were complemented with an analysis of variance. Also in this situation the validation was based on the bootstrapping validation with $N = 1000$ bootstrap samples.

This validation was repeated for all variables and then the same reasoning was applied to some combinations of two and three variables. These combinations were formed considering that those risk factors belonged to at least two of the individual models that integrate the combination scheme.

3.5 Incorporation of Clinical Knowledge

An important limitation of the current risk assessment models is the inability to incorporate additional clinical knowledge (clinical expertise)[90].

For a naïve Bayes model the incorporation of a new risk factor is a very straightforward process, since it represents just one more attribute X_i. As a result, the incorporation of a new attribute represents one more conditional probabilities table that has to be considered by the Bayesian inference mechanism. This possibility is a significant potential advantage of the proposed Bayesian approach.

The selection of the additional clinical knowledge and the consequent definition of the relationship between the risk factor categories and the risk levels must be defined by the physicians.

The incorporation of Body Mass Index (BMI) in Bayesian individual models that were derived based on current risk assessment tools was assessed. Actually, the cardiologists that collaborated in this thesis identified the BMI as an important risk factor that should be included in the CVD risk assessment. Moreover, there are several recent research works that intend to create new CVD risk scores to

[89] Similarly to the previous validation scenario and in relation to Voting model, if the missing variables were continuous the replacement was done based on the respective mean values, in the case of Boolean variables their value was successively replaced by 0 and 1 values.

[90] The cardiologists that collaborated in this research found very interesting the possibility of incorporating additional knowledge (new risk factor) in the risk assessment model.

incorporate the BMI, which demonstrate the importance of this risk factor *(Wormser, 2011) (Dudina, 2011).* The validation procedure is presented in Figure 3.7.

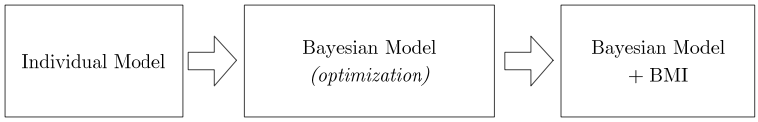


Figure 3.7 - Validation strategy to the BMI incorporation.

The main steps of this procedure are:

- Derivation of each individual model. A training dataset was submitted to each individual risk assessment tool, originating in the required parameters to define the individual Bayesian models (naïve Bayes);
- An optimization was performed to adjust the behavior of individual Bayesian models to the respective risk assessment tool. The main goal of this phase was to assure that the Bayesian model reproduces the behavior of original statistical model as accurately as possible;
- The last phase was the integration of the new risk factor. This step was performed through the concatenation of the conditional probabilities table of the model with the conditional probabilities table of the new risk factor.

The validation was performed comparing the individual models' performance before and after the incorporation of the new risk factor (BMI). This assessment was made through bootstrapping validation based on a real patient testing data set.

3.6 Personalization based on Grouping of Patients

As mentioned this thesis aims to minimize some of the identified weaknesses of the current CVD risk assessment tools. One of the main goals is the improvement of the performance (SE/SP) of risk assessment. However, there were some test cases where the combination methodology did not achieve a significant improvement of the performance, namely of the specificity's value. Actually, the contributions of individual models that present low performances (low sensitivity and/or low specificity) may impose an additional difficulty to the global model to assure a correct risk prediction. Additionally, the implemented optimization procedure, *multiobjective optimization*, where a tradeoff between objectives (maximize sensitivity/maximize specificity) must be found, may not be able to correct this flaw.

In this context, an alternative approach based on a personalization strategy is proposed. This methodology relies on the evidence that risk assessment tools perform differently among different populations. This variation of performance indicates that a specific risk assessment tool may have a good performance within a given group of patients and performs poorly within other groups.

The main hypothesis that supports this methodology can be stated as:

- If the patients are properly grouped (clustered) it would be possible to find the best classifier for each group.

The methodology (Figure 3.2) is composed of two main phases: *i*) grouping of patients; *ii*) selection of risk tools.

3.6.1 Grouping of Patients

The grouping of patients phase involves two main steps: *i*) dimension reduction; *ii*) clustering.

As mentioned, the proposed personalization strategy relies on the creation of groups of patients. However, the heterogeneity of risk factors (quantitative data, qualitative data, binary data) that usually characterize a specific patient, along with their high dimensionality (number of risk factors) constrain the derivation of those groups. Therefore, the reduction of dimensionality is implemented in order to facilitate/improve the clustering process.

The second step consists of a clustering procedure, where groups of patients are created based on the information obtained through the dimension reduction procedure.

Dimension Reduction

Section 2.4.1 presents an overview of the main linear/nonlinear methods to implement the reduction of dimensionality. However a different approach is followed in this work. The reduction of dimensionality process is supported on the individual risk assessment tools (non-linear mapping). In effect, this approach seems very appropriate in this particular problem as these tools were developed to classify patients that are characterized by a set of heterogeneous risk factors. Additionally, this non-linear mapping allows the uniformization of each patient's data.

All instances $\mathbf{x}_i = [x_1^i ... x_p^i]^T \in \mathbf{X}_{p \times N}$, that correspond to the N patients are

mapped into $\mathbf{y}_i \in \mathbf{Y}_{Q\times \mathrm{N}}$, $i = 1,...,N$ where y_q^i denotes the output of tool q to classify the patient i (e.g. $\mathbf{y}_i = [y_R^i \; y_P^i \; y_T^i \;]$[91]).

Assuming that a risk assessment tool q considers J risk factors (subset of the p risk factors), an instance $\mathbf{x}_i^q$ (containing the J risk factors (values) of patient i) is applied to the q tool in order to obtain the respective $y_q^i \in \mathbf{y}_i$. All the y_q^i should be normalized into the interval $[0,1]$.

Clustering

This phase is responsible for the creation of the patient groups. Basically, using the proposed approach, patients are grouped based on the outputs of the risk tools instead on the initial risk factors. Let $\mathbf{Y}_{Q\times N}$ represent a set of N patients, the goal is to apply a clustering algorithm to $\mathbf{Y}_{Q\times N}$ in order to create K disjoint groups (clusters) $G = \{G_1,...,G_K\}$ of patients with similar characteristics.

The clustering process should assume that the dimension of the clusters must be defined considering the concept that supports the methodology, i.e. if the cluster is too big it may not provide a differentiation among the performance of the several risk assessment tools otherwise if the cluster is to small it will be impossible to apply the concept of patient grouping.

As a result of this step, the global set of patients $D = \{(\mathbf{x}_1, c_1),...,(\mathbf{x}_N, c_N)\}$ is clustered in K clusters G_i which originate the separation of the patients as presented in (3.17):

$$D = \bigcup_{i=1}^{K} D_i \tag{3.17}$$

It is important to emphasize that a given patient belongs only to one cluster G_i. The initial clusters are created through the subtractive clustering technique[92] that was previously introduced.

---

[91] In this thesis the strategy validation was performed with g**R**ace, **P**ursuit and **T**imi risk assessment tools.

[92] Other algorithms can be used for this initial separation of the patients, e.g. k-means.

3.6.2 Identification of Risk Tools

The second phase concerns the selection of the most suitable tool to classify patients from a given cluster. The performance of the several individual tools is assessed within each group of patients (created in the previous phase). This allows that each cluster be assigned the tool that presents the best performance. The final classification of a particular patient that belongs to a given cluster corresponds to the classification of the individual tool that has the best performance with patients from that cluster.

Tools Assessment

Each one of the considered individual risk assessment tools is tested within each cluster. Thus, each y_q^i is converted to a risk class c_q^i according to the original specifications of each tool. Then for each patient i of each cluster G_k , $k = 1,...,K$ the output (class) of each tool q is compared with the real data (occurrence of an event) within a given period of time.

This assessment allows computing the sensitivity and specificity of the risk prediction achieved by each tool.

Selection

The final classification of a patient is based on the selection of the most suitable risk tool for its classification.

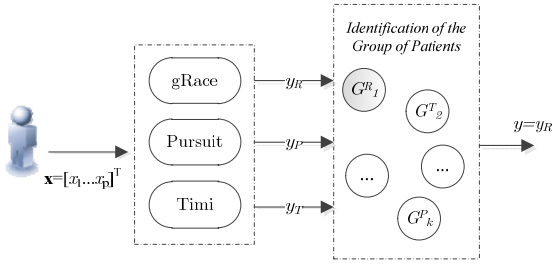


Figure 3.8 - Classification[93]

The classification process may be depicted as follows: *i*) the different risk assessment tools assess the risk of a new patient i based on $\mathbf{x}_i$ in order to obtain $\mathbf{y}_i$;

[93] G_k^q denotes that tool q has the best performance on cluster G_k

ii) the cluster G_k that the patient i belongs to is identified based on $\mathbf{y}_i$; *iii*) the best tool q to classify patients from G_k is selected; *iv*) the final classification is provided by that tool.

The criteria to select the best tool q to classify patients from a cluster G_k are defined based on the values of G_{mean}, SE, SP obtained by that tool q in that specific cluster G_k. Figure 3.9 presents the selection algorithm:

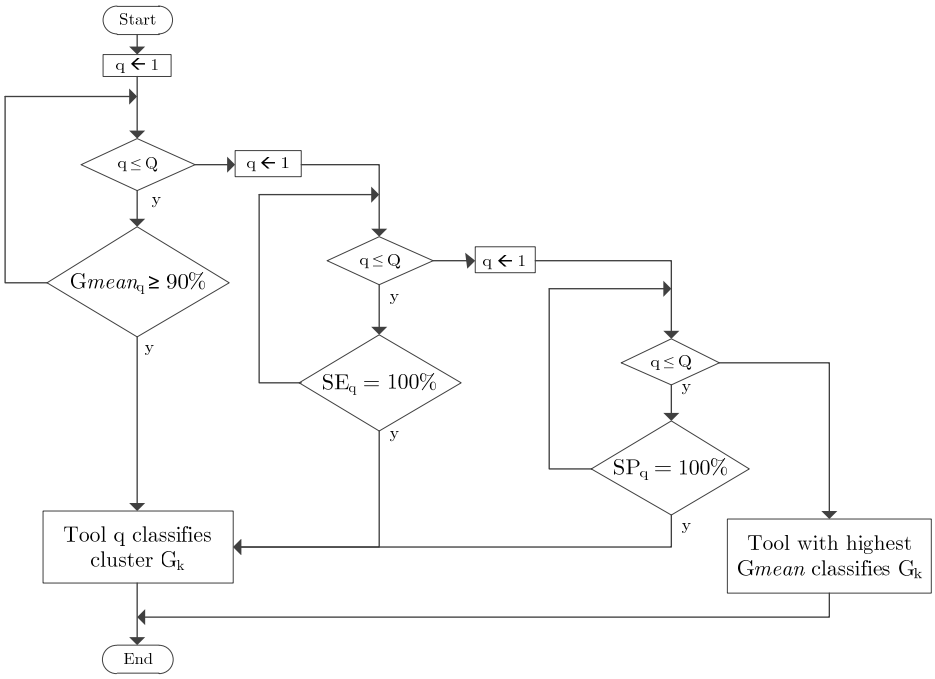


Figure 3.9 - Selection algorithm. [94]

3.7 Validation of the Personalization Methodology

Similarly to the validation procedure depicted in Section 3.4, two validation scenarios were implemented: *i*) simulation - theoretical individual models; *ii*) tools applied in clinical practice.

In both scenarios a dataset was applied to the individual models/tools. This procedure allowed the dimensionality reduction needed for grouping patients. The performance of each classifier was assessed in each cluster in order to identify the best

[94] Q represents the total number of tested tools. The selection procedure is applied to the K clusters.

model/tool to classify the patients that belong to a specific cluster. According to the validation scenario, different testing datasets were applied to evaluate the performance of this personalization approach.

Simulation - Theoretical Individual Models

The first scenario was developed based on models derived through Cox regression directly from the TEN-HMS dataset[95]. A data set was generated $\Upsilon = \{\mathbf{x}_1,...,\mathbf{x}_N\}$ and it was applied to the derived individual models M_i in order to group patients as well as to obtain the respective complete labelled datasets $D_i = \{(\mathbf{x}_1, c_1^i),...,(\mathbf{x}_N, c_N^i)\}$. The same dataset Υ was applied to the model that comprises all the considered variables (complete Cox model) to obtain the *true data* $D_T = \{(\mathbf{x}_1, c_1^T),...,(\mathbf{x}_N, c_N^T)\}$.

Each D_i was compared with D_T for patients of each group that resulted from the grouping of patients phase (Figure 3.2). This allowed the assessment of sensitivity and specificity obtained by each individual model in each individual group which permitted the selection of a specific model to a given cluster (Figure 3.9).

A testing dataset $O = \{\mathbf{x}_1,...,\mathbf{x}_M\}$ was generated to test the proposed approach. The testing dataset O was applied to the complete Cox model (*true data*) as well as to each individual model M_i. Then, each patient that belongs to O was assigned to the respective group and the proposed methodology (Figure 3.8) was implemented. The global assessment was performed through the comparison of the output of the personalization strategy with the *true data* as well as with the comparison of each individual model's output with the *true data*. The complete testing procedure was repeated $n = 30$ times for enhancing the statistical significance of the obtained results.

Tools Applied in Clinical Practice

The latter validation scenario was comprised of three risk assessment tools[96] that are currently applied in the daily clinical practice. The *true data* was directly obtained from the Santa Cruz hospital dataset that is described in Section 4.2.2. The outputs of the three risk assessment tools were computed considering all the patients that belong to the dataset. Patients were grouped and the SE/SP of each risk

[95] This derivation procedure is detailed in Section 3.4.1.

[96] GRACE, PURSUIT; TIMI.

assessment tool were assessed in each group of patients. Due to limitations in the available dataset[97], bootstrapping validation $N_B = 1000$ was adopted to generate the required testing datasets to reinforce the reliability of the obtained results.

3.8 Conclusions

The methodologies and respective algorithms that were developed and implemented in this thesis were described in this Chapter.

Figure 3.1 presents the combination of individual risk assessment tools methodology that permits to avoid some of the identified weaknesses of the current risk assessment tools. This methodology is composed of three phases: *i*) common representation of individual models. The naïve Bayes is the selected classifier to implement this uniform representation; *ii*) individual models' combination. The models are combined based on a new combination scheme that is designated as individual model parameters' weighted average. This models' combination is included in the model parameter/data fusion as it combines the individual model's parameters directly; *iii*) validation. The availability of data highly influenced the validation options, originating the two validation scenarios that are detailed in Sections 3.4.1 and 3.4.2. Closely related with this approach is the incorporation of additional clinical knowledge/new risk factor, as it can be implemented through a direct application of the combination methodology.

The previous approach circumvents the majority of the identified weaknesses of the current risk assessment tools, although it does not assure higher sensitivity/specificity than the current tools in all situations. This is a flaw of the methodology that must be further investigated.

To minimize this difficulty, an additional approach (Figure 3.2) based on a personalization concept is proposed. This new methodology assumes that if the patients are properly grouped (clustered) it would be possible to find the best classifier for each group. Therefore, its main goal is to select the most suitable classifier considering the characteristics of a specific patient. It is important to emphasize that this last methodology was presented as an alternative strategy that intends to improve the performance (SE/SP) of the risk assessment. However, the

[97] Severely imbalanced dataset (low events rate).

ongoing research can merge the two developed strategies (Figure 3.1; Figure 3.2) obtaining the main advantages of both approaches.

4. Results

4.1 Introduction

In this chapter the main validation results are presented according to the validation strategies defined in Chapter 3. The Figure 4.1 provides a global perspective of the issues addressed in this chapter.

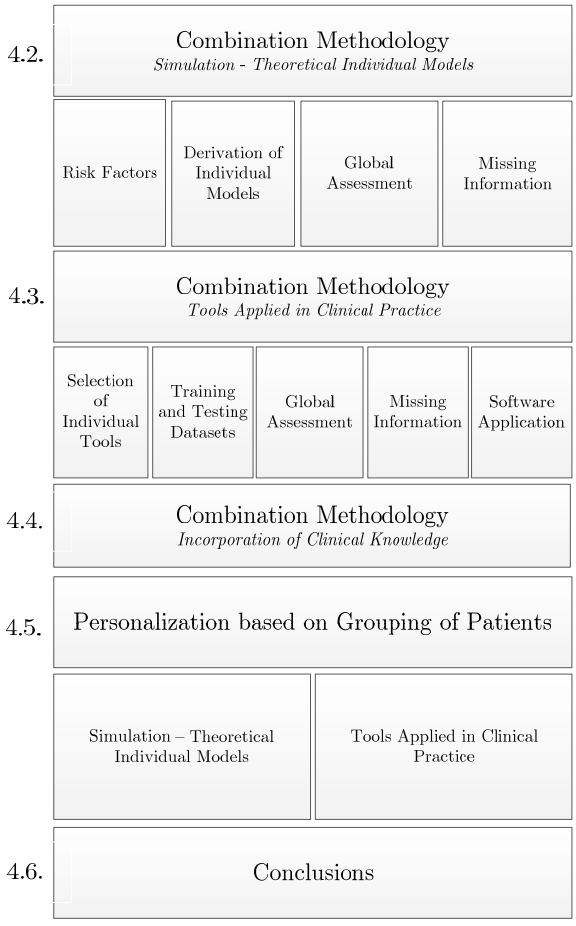


Figure 4.1 - Structure of chapter 4.

In Section 4.2 the results related with validation of the combination methodology applied to the simulated individual models are explored. The available dataset (TEN-HMS dataset) is detailed along with the individual simulated models (theoretical) that were derived from it. The combination scheme's performance is assessed as well as its capability to cope with missing risk factors. This validation procedure involves patients with heart failure.

The validation of the combination methodology when applied to tools that are currently available in the daily clinical practice is described in Section 4.3. The available testing datasets (Leiria-Pombal Hospital Centre/Portugal, Santa Cruz Hospital/Portugal) are detailed. Several validation issues are explored, such as: *i*) performance of individual risk assessment tools, *ii*) assessment of Bayesian global model; *iii*) optimization; *iv*) ability to deal with missing risk factors. This validation scenario is applied to coronary artery disease patients.

Section 4.4 addresses the incorporation of additional clinical knowledge, namely the validation of the integration of BMI risk factor into the current risk assessment models. This incorporation configures a direct application of the combination methodology proposed in this work. To evaluate the eventual influence of the BMI the individual risk tools' performance before and after its incorporation was assessed.

The personalization strategy based on grouping of patients was also validated according to two different scenarios: *i*) simulation – theoretical individual models; *ii*) tools applied in the clinical practice. The performance of the CVD risk assessment obtained through the personalization approach was compared with the one achieved by the individual risk assessment models/tools. These results are presented in Section 4.5.

4.2 Simulation – Theoretical Individual Models

This validation procedure was developed based on simulated models derived from a real heart failure patient's dataset (TEN-HMS dataset). The validation strategy was significantly influenced by some specific limitations of the available dataset.

The assessment of the models' performance was based on three different metrics: *i*) accuracy; *ii*) sensitivity; *iii*) specificity. Different statistical significance tests were applied to reinforce the obtained results: *i*) Student's t-test for comparison of means

between two unpaired groups. In some specific situations the results of this test were compared to the ones obtained with Mann Whitney U test (non-parametric test); *ii*) Levene's test was applied to compare variances between two unpaired groups; *iii*) One-way ANOVA was implemented to compare the metrics' mean value obtained with more than two models.

4.2.1 Risk Factors and Complete Cox model

Variable	Mean ± std
Age (years)	67.1 ± 11.7
Height (cm)	170.92 ± 9.55
Weight (kg)	76.65 ± 16.82
Gender: Male /Female	325 (77%) / 97 (23%)
SBP (mmHg)	114.36 ± 19.26
DBP (mmHg)	69.31 ± 11.31
Ejection Fraction (%)	25.09 ± 7.6
Hemoglobin (g/dl)	13.1 ± 2.27
White Cell Count	8.362 ± 2.83
Sodium (mmol/l)	135.55 ± 16.91
Urea (mg/dl)	11.06 ± 6.78
Creatinine (mg/dl)	135.3 ± 52.97
AbnormalSinusRhythm (0/1)	143(33%) / 283 (67%)
AF (0/1)	329 (77%) / 97 (23%)
First MI (0/1)	101 (23%) / 325 (77%)
ValvularDisease (0/1)	228 (53%) / 198 (47%)
ChronicAF (0/1)	270 (63%) / 156 (37%)
Pacemaker (0/1)	316 (74%) / 110 (26%)
Defribillator (0/1)	371 (87%) / 55 (13%)
RenalFailure (0/1)	392 (92%) / 34 (8%)
DiabetesInsulin (0/1)	304 (71%) / 122 (29%)
Cancer (0/1)	350 (82%) / 76 (18%)
NYHA: I/II/III/IV	17 (4%) /59 (14%) /91 (22%)/ 255 (60%)
Diuretics (0/1)	12 (2%) / 414 (98%)
KsparingDiuretics (0/1)	309 (72%) / 117 (28%)
Bblockers (0/1)	194 (44%) / 232 (56%)
ACEinhibitor (0/1)	72 (16%) / 354 (84%)
Statin (0/1)	333 (78%) / 93 (22%)
Allopurinol (0/1)	384 (90%) / 42 (10%)
Stroke (0/1)	422 (99%) / 4 (1%)
Angiography (0/1)	418 (98%) / 8 (2%)
Survival (days)	312.78 ± 103.99

Table 4.1 - Clinical characteristics of patients that integrate the TEN-HMS.

The dataset (Table 4.1) was obtained from the TEN-HMS dataset[98] and contains data from $N = 426$ patients. It consists of 31 variables, 12 of which are continuous. The 1 year's endpoint rate of the available dataset is 29.5% which corresponds to 126 deaths.

Identification of variables (risk factors)

The set of variables chosen for the derivation of individual models was based on the relevant tools for assessing the one year death risk. However, this selection was also guided by the availability of data.

Table 4.2 shows the variables as well as the way that they are considered in this work.

Raw variable	Variables	Type
Years	Age	discrete [35..90]
NYHA	NYHA Class III/IV	Boolean
VHD	Valvular heart disease	Boolean
Diabetes	Diabetes	Boolean
Creatinine	MSKD	Boolean > 2 mg/dL
Patients on dialysis	MSKD	Boolean
Transplant or Uremia	MSKD	Boolean
Metastatic cancer	Cancer	Boolean
Number of cancers	Cancer	Boolean >=2
SBP	Hypertension	Boolean>140 mmHg
DBP	Hypertension	Boolean>90 mmHg
LVEF	LVEF	Boolean <20%
AF	Atrial fibrillation	Boolean
Hemoglobin	Anemia	Boolean<11 g/dL
Bblockers	No Bblockers	Boolean
ACE	No ACE	Boolean

MSKD – Moderate/Severe Kidney Dysfunction; **LVEF** - Left Ventricular Ejection Fraction; **ACE**-Angiotension Converting Enzyme

Table 4.2 - Identification of variables.

Among these twelve variables ($p = 12$), eleven are Boolean. Age is continuous and it was discretized as follows: [55- less than 60; 60- less than 65; 65-less than 70; 70- less than 75; 75-less than 80; 80-less than 85; 85-less than 90].

One year was the time for the risk assessment of the developed models. Each model's output (risk) had two possible values (low/intermediate risk $\leq 30\%$, high risk $> 30\%$). This cut-off value can be easily adjusted.

[98] This dataset was made available by Castle Hill Hospital, Hull, UK.

Complete Cox model

A complete model based on all variables $(p = 12)$ and all patients $(N = 426)$ was directly derived from the available dataset, through Cox regression. When validated in that dataset it presented a sensitivity of 73.1% and a specificity of 54%, which originated an AUC of 0.650. The main function of this model was to support the validation by means of: *i*) its comparison with individual models performance and *ii*) its comparison with the result of the combination approach.

4.2.2 Derivation of Individual Models

Definition of individual models

Based on the algorithm described in Section 3.4.1, twenty-two subsets of the original dataset, i.e., twenty-two individual models were defined $(M = 22)$.

Model	hypert	lvef	anemia	Af	age	valvular	renal	diabetes	cancer	nyha	Bblocker	ace
M1	▪		▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
M2		▪			▪	▪	▪	▪	▪			
M3	▪	▪	▪	▪			▪	▪	▪	▪	▪	▪
M4	▪	▪	▪	▪	▪	▪	▪		▪	▪		
M5	▪	▪	▪	▪						▪	▪	▪
M6	▪					▪	▪		▪	▪		
M7		▪	▪	▪	▪			▪			▪	▪
M8		▪	▪	▪	▪	▪			▪		▪	▪
M9	▪			▪			▪	▪	▪	▪		
M10		▪	▪				▪	▪	▪			
M11	▪	▪	▪	▪			▪	▪	▪	▪	▪	▪
M12		▪			▪	▪	▪	▪		▪		
M13	▪						▪	▪		▪		▪
M14		▪	▪	▪	▪	▪			▪		▪	
M15	▪					▪	▪	▪		▪		▪
M16	▪	▪				▪	▪	▪		▪	▪	▪
M17		▪	▪	▪	▪	▪	▪		▪	▪	▪	▪
M18		▪		▪	▪			▪			▪	▪
M19	▪		▪			▪	▪		▪	▪		
M20	▪	▪	▪	▪	▪		▪	▪	▪	▪		
M21	▪		▪	▪	▪	▪	▪	▪	▪	▪		
M22	▪	▪	▪	▪		▪	▪	▪	▪	▪	▪	▪

Table 4.3 - Composition of the individual models.

Table 4.3 presents the subset of the risk factors considered for each one of these individual models. For instance, M6 was composed of only 5 risk factors (*hypert*, *valvular*, *renal*, *cancer* and *nyha*).

According to the proposed algorithm, depicted in *(Kutner, 2004)*, ten samples ($S = 10$) of 142 patients were used to create ten distinct regression models for each one of the $M = 22$ subsets of variables (individual model) presented in Table 4.3. The available dataset composed of 426 patients was divided, as shown in Table 4.4.

According to this distribution, e.g. S1 contains patients from the 1st to the 141st patient, S2 from 142nd to the 282nd patient.

Patients	Samples									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
1-47	▪							▪		▪
48-94	▪			▪	▪				▪	▪
95-141	▪			▪			▪			
142-188		▪			▪				▪	
189-235		▪				▪				
236 - 282		▪		▪	▪		▪			
283 - 329			▪			▪		▪		
330 - 376			▪					▪	▪	
377- 426			▪			▪	▪			▪

Symbol (▪) means that patients belong to a specific sample.

Table 4.4 - Samples definition.

Selection of individual models

As a result, by considering ten distinct parameterizations ($S = 10$) for each individual model ($M = 22$) a total of 220 Cox regression models were derived[99].

Table 4.5 presents the accuracy values of each individual model M_i, namely the mean and variance of accuracy considering the ten instances.

The models marked with symbol (▪) in Table 4.5 were rejected as they had a variance higher than 75[100]. The remaining models ($M = 13$) had potential to be combined.

[99] The calculation of regression coefficients and the survival function for each model were developed with *Mathworks Matlab /Statistics Toolbox.*

[100] This limit was imposed regarding the minimum number of models required to perform a number of combinations that assured the statistical significance of the obtained results.

	Model	S	μ	σ^2
	M1	10	77.21	18.30
	M2	10	82.32	5.04
■	M3	10	62.63	135.06
	M4	10	80.33	18.76
■	M5	10	61.36	146.29
■	M6	10	55.11	123.33
	M7	10	77.60	23.59
	M8	10	78.06	18.89
	M9	10	64.82	65.01
■	M10	10	66.45	92.50
■	M11	10	62.63	135.06
	M12	10	82.19	9.65
■	M13	10	53.87	270.89
	M14	10	81.66	7.67
■	M15	10	54.22	203.41
■	M16	10	62.15	100.35
	M17	10	79.04	9.57
	M18	10	78.73	23.90
■	M19	10	61.38	92.08
	M20	10	80.90	4.89
	M21	10	80.16	16.56
	M22	10	64.88	73.65

S - Number of samples; μ - Mean; σ^2 - Variance

Table 4.5 - Individual models' accuracy.

Based on the selected models $(M = 13)$, 29 possible test cases (combinations of two models[101]) were considered, $T_i, \quad i = 1,..,29$.

The definition of each test case had to verify the following criterion: *each variable belongs at least to one individual model*. Using this procedure, in each test case the two models covered all the twelve available variables.

Table 4.6 depicts these twenty-nine test cases. For example, test T1 combined the individual models M1 and M2, T2 combined M1 and M4, etc. The number of tests was the minimum required to assure the statistical significance of the obtained results.

[101] As mentioned, a reduced number of individual models is usually combined in the clinical practice. For this reason the combination strategy was validated considering only two individual models.

	M1	M2	M4	M7	M8	M9	M12	M14	M17	M18	M20	M21	M22
M1		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10		T11
M2													T12
M4				T13						T14			T15
M7												T16	T17
M8											T18	T19	T20
M9													
M12													T21
M14													T22
M17											T23	T24	T25
M18												T26	T27
M20													T28
M21													T29
M22													

Table 4.6 - Combination tests.

Training dataset

The derivation of each individual Bayesian classifier to replicate the behavior of each individual regression model (Table 4.5) required the respective parameter learning. Then a set of instances $\Upsilon = \{\mathbf{x}_1,...,\mathbf{x}_N\}$ is needed to obtain the required training dataset $D = \{(\mathbf{x}_1, c_1),...,(\mathbf{x}_N, c_N)\}$. Each instance $\mathbf{x}_i$; $i = 1,...,N$ was applied to each individual regression model j and the respective c_i^j was obtained. Prior and conditional probabilities for each individual Bayesian classifier were derived based on the resulting dataset $D^j = \{(\mathbf{x}_1, c_1^j),...,(\mathbf{x}_N, c_N^j)\}$

Continuous variable (age) was generated from normal distribution (67.0 ± 11.7) and discretized through the method EWD. The remaining binary variables were randomly generated. The training dataset was composed of $N = 1000$ instances.

The set of instances to generate the testing dataset ($N = 1000$) was obtained based on an identical procedure.

4.2.3 Global Assessment

The validation of the proposed combination strategy aimed to assess the performance of individual models (when considered independently), compared with the one obtained by the global model that resulted from the respective combination.

Table 4.7 presents the performance of the individual models as well as the model that resulted from their combination in the 29 test cases. Combination performance was assessed before and after the genetic algorithm operation.

All the formulas were calculated taking into account the complete testing dataset[102] obtained through the complete Cox model.

Test	Acc_{IM}	SE_{IM}	SP_{IM}	Acc_{BG}	SE_{BG}	SP_{BG}	Acc_{BGAO}	SE_{BGAO}	SP_{BGAO}
T1	87.1	92.3	59.4	90	100	37.1	93.4	97.0	74.0
T2	92.4	95	78.9	90.8	100	42.1	92.2	95.5	74.8
T3	89.8	97.5	49.0	86.9	100	17.6	93.2	95.9	78.6
T4	86.9	97.9	44.0	85.4	100	18.0	93.4	98.3	67.3
T5	80.4	81.4	74.8	75.9	87.3	15.7	84.1	99.8	30.0
T6	85.5	90.0	60.9	91	99.5	45.9	92.4	96.5	70.4
T7	89.2	97.4	38.7	86.7	100	16.3	87.0	100	18.2
T8	90.3	97.3	53.1	86.8	100	16.9	87.1	100	18.7
T9	88.5	96.6	43.7	87.1	100	18.9	88.9	99.8	30.0
T10	92.2	94.4	80.8	91.3	100	45.2	91.3	100	45.2
T11	88.8	90.7	78.2	93.1	99.4	59.7	94.4	96.5	83.0
T12	82.8	96.9	60.4	91.0	97.3	57.2	91.0	97.3	57.2
T13	89.5	94.4	50.2	87.7	100	22.6	91.8	99.6	50.3
T14	87.9	95.9	45.2	87.4	100	20.7	88.4	99.8	30.0
T15	88.1	89.6	79.9	92.6	98.6	60.3	92.6	98.6	60.3
T16	87.8	94.8	50.6	88.7	100	30	88.7	100	30.0
T17	85.5	92.1	49.2	89.2	99.8	32.7	90.7	97.8	52.8
T18	86.5	96.2	47.1	87.9	100	24.0	88.5	100	25.3
T19	87.4	95.3	44.5	88.0	100	24.5	90.1	99.2	42.0
T20	85.1	92.6	44.5	87.9	99.5	26.4	88.4	99.5	30.0
T21	81.2	84.7	61.8	91.5	96.6	64.1	91.9	96.7	66.1
T22	84.9	92.1	46.0	88.1	99.4	28.3	88.9	88.9	98.8
T23	89.4	95.6	56.5	88.5	100	27.6	88.5	100	27.6
T24	88.3	94.7	54.3	88.5	100	27.8	88.5	100	27.8
T25	86.0	91.9	54.0	89.5	99.8	34.5	90.9	98.6	49.7
T26	86.5	94.3	44.9	88.1	99.8	25.7	89.7	98.8	41.5
T27	84.2	91.6	44.6	87.4	98.9	26.4	87.4	97.3	34.5
T28	87.9	89.0	87.7	93.6	98.2	69.0	93.6	98.2	69.2
T29	88.1	88.1	79.5	92.5	97.5	66.0	96.0	98.6	81.2

Acc_{IM} – Accuracy of individual models ; **SE_{IM-}** – Sensitivity of individual models; **SP_{IM-}** – Specificity of individual models; **Acc_{BG}** – Accuracy of the combination scheme; **SE_{BG}** - Sensitivity of the combination scheme; **SP_{BG}** - Specificity of the combination scheme; **Acc_{BGAO}**– Accuracy of the combination scheme after optimization; **SE_{BGAO}** - Sensitivity of the combination scheme after optmization; **SP_{BGAO}** - Specificity of the combination scheme after optimization

Table 4.7 - Assessment of models' performance.

[102] The testing dataset was generated through an identical approach to the one adopted for the training dataset.

Table 4.8 contains some descriptive statistics of the formulas assessed:

	Acc_{IM}	SE_{IM}	SP_{IM}
mean	87.2 *(86.1;88.2)*	*93.1 (91.6;94.6)*	*57.3 (51.8;61.7)*
std.	2.8	3.9	14.3

a) Individual models

	Acc_{BG}	SE_{BG}	SP_{BG}
mean	88.7 *(87.4;89.9)*	*99.0 (98.0;99.9)*	*34.5 (28.7;40.9)*
std.	3.3	2.43	16.7

b) Bayesian Global model before optimization

	Acc_{BGAO}	SE_{BGAO}	SP_{BG}
mean	90.4 *(89.4;91.4)*	*98.2 (97.3;99.0)*	*50.5 (42.0;58.9)*
std.	2.7	2.3	15.7

(-;-)= 95% CI

c) Bayesian Global model after optimization

Table 4.8 – Assessed metrics statistics (% values).

Student's t-test/Levene's test were performed to obtain more reliable conclusions about the data presented in Table 4.7.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	7.71	0.007	6.0	56	0.000	5.1	0.84	3.41	6.78
ByGAO *vs* ByIM				6.0	46	0.000	5.1	0.84	3.41	6.78
SP	ea	8.56	0.005	-1.38	56	0.171	-6.8	4.92	-16.68	3.03
ByGAO *vs* ByIM				-1.38	47	0.172	-6.8	4.92	-16.68	3.03
Acc	ea	0.048	0.828	4.55	56	0.000	3.2	0.718	1.83	4.70
ByGAO *vs* ByIM				4.55	56	0.000	3.2	0.718	1.83	4.70

Levene's Test: *ea:* equal variances assumed; F: F statistics value; Sig.: if *p-value* <0.05 null hypothesis must be rejected; **t-test:** t: t-test statistics value; df: degrees of freedom; Sig.(2-tailed): if *p-value* <0.05 null hypothesis should be rejected; **ByGAO**– Bayesian Global Model After Opt.*;* **ByIM**– Bayesian Individual Models,

Table 4.9 - Bayesian after optimization vs. individual models[103].

[103] IBM SPSS statistical software's output.

In both tests, if the *p-value* (significance value) is lower than 0.05, then there is strong evidence against the null hypothesis[104] (null hypothesis should be rejected) which means that the equality of means/variances should not be assumed.

- There was strong evidence against the equality of means of sensitivity and accuracy between Bayesian global model after the optimization and the individual models. Both values were higher in the global model (*SE* 98.2; 95%*CI* : 97.3 *to* 99.0 / Acc 90.4; 95%*CI* : 89.4 *to* 91.4) than in the individual models (*SE* 93.1; 95%*CI* : 91.6 *to* 94.6 / Acc 87.2; 95%*CI* : 86.1 *to* 88.2). This strong evidence against the equality of means indicated that differences between the two mean values cannot be exclusively attributed to sample error;
- In contrast, the specificity value was higher in the individual models (*SP* 57.3; *CI* : 51.9 *to* 62.7) than in the global model (*SP* 50.5; *CI* : 42.0 *to* 58.9). However in this case the t-test is not conclusive as the null hypothesis (equality of means) should not be rejected (*p-value* = 0.171).

The Mann Whitney U test[105] was adopted to reinforce these conclusions:

	SE	SP	Acc
	ByGAO *vs* ByIM	ByGAO *vs* ByIM	ByGAO *vs* ByIM
Mann-Whitney U	71.500	327.000	161.50
Wilcoxon W	506.500	762.000	596.50
Z	-5.433	-1.455	-4.02
Asymp. Sig. (2-tailed)	0.000	0.146	0.000

Table 4.10 - Bayesian global model after opt. vs. individual models (Mann- Whitney U Test).

This test confirms that the null hypothesis should be rejected in the sensitivity and accuracy metrics. These results strengthen the conclusions derived with the parametric test.

Table 4.11 presents the results of the comparison of means between the global model resultant from the combination scheme before and after the optimization procedure.

---

[104] (Student's t-test) H_0: The mean values of two datasets are equal *vs.* H_1: The mean values of two datasets are not equal; (Lenene's test) H_0: The variances of two datasets are equal (homogeneity of variances) *vs.* H_1: The variances of two datasets are not equal.

[105] H_0: The medians between the two classifiers are equal *vs.* H_1: The medians of the two datasets are not equal.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	0.322	0.573	-1.3	56	0.199	-0.81	0.620	-2.04	0.44
ByGAO *vs* ByG				-1.3	55	0.199	-0.81	0.620	-2.04	0.44
SP	ea	3.936	0.052	3.08	56	0.003	15.97	5.185	5.58	26.36
ByGAO *vs* ByG				3.08	51	0.003	15.97	5.185	5.57	26.38
Acc	ea	0.009	0.926	2.19	56	0.032	1.72	0.784	0.14	3.29
ByGAO *vs* ByG				2.19	53	0.033	1.72	0.784	0.14	3.29

ByGAO – Bayesian global model after optimization; ***ByG*** – Bayesian global model before optimization,

Table 4.11 - Bayesian after optimization vs Bayesian before optimization.

It is possible to conclude that:

- Sensitivity value did not improve with the optimization. The equality of means should not be rejected;
- Specificity improved $(mean\ diff = 15.97;\ pvalue = 0.003)$. In this case, the optimization procedure contributed significantly for the improvement of the performance of the global classifier;
- As a result of the growth of specificity, accuracy also improved. According to the t-test, the hypothesis H_0 should be rejected.

Adopting the same procedure, Table 4.14 presents the results extracted with the non-parametric test:

	SE	SP	Acc
	ByGAO *vs* ByG	ByGAO *vs* ByG	ByGAO *vs* ByG
Mann-Whitney U	266.000	224.500	276.500
Wilcoxon W	701.000	659.500	711.500
Z	-2.473	-3.050	-2.241
Asymp. Sig. (2-tailed)	0.013	0.002	0.025

Table 4.12 - Bayesian after opt. vs. Bayesian before opt. (Mann- Whitney U test).

Based on Table 4.12 the null hypothesis should be rejected for the three assessed metrics. These conclusions partially confirmed those obtained with the parametric tests. In fact, considering the sensitivity values the Student's t-test did not reject the equality of means while the Mann-Whitney U test rejected the equality of medians which suggests a significant difference between the classifiers. The two statistical

tests presented contradictory results however analyzing the data in Table 4.8 it seems that the optimization did not have a significant impact in the sensitivity value.

The analysis of variance provides a global picture of the relationships between the different classifiers presented in the Table 4.7.

	Levene Statistic	df1	df2	Sig.
SE	6.258	2	84	0.003
SP	4.690	2	84	0.012
Acc	0.017	2	84	0.983

Table 4.13 - Test of homogeneity of variances.

The homogeneity of variances is an important assumption of ANOVA. Although it can not be assumed in relation to sensitivity and specificity values (Table 4.13). To circumvent this issue, an analysis of variance was performed with a post-hoc method[106] based on the inequality of variances (e.g Tamhane's T2 method, Dunnett's T3, Dunnett's C, etc.) *(IBM , 2010).*

		Sum of Squares	df	Mean Square	F	Sig.
SE	Between Groups	595.008	2	297.504	33.607	0.000
	Within Groups	743.597	84	8.852		
	Total	1338.604	86			
SP	Between Groups	7942.491	2	3971.246	12.100	0.000
	Within Groups	27568.166	84	328.192		
	Total	35510.657	86			
Acc	Between Groups	155.093	2	77.546	9.060	0.000
	Within Groups	718.958	84	8.559		
	Total	874.051	86			

Table 4.14 – Comparison of classifiers [ANOVA].

The ANOVA results (Table 4.14) show that the null hypothesis among the several classifiers should be rejected which indicates that there are differences between the classifiers' performance.

[106] 'post-hoc' tests are applied for further explanation after a significant effect has been found *(Hilton, 2006).*

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
SE	1.00	2.00	-0.806	0.620	0.486	-2.334	0.720
		3.00	5.100	0.843	0.000	3.009	7.190
	2.00	1.00	0.806	0.620	0.486	-0.720	2.334
		3.00	5.906	0.857	0.000	3.782	8.030
	3.00	1.00	-5.100	0.843	0.000	-7.190	-3.009
		2.00	-5.906	0.857	0.000	-8.030	-3.782
SP	1.00	2.00	15.975	5.185	0.010	3.182	28.769
		3.00	-6.824	4.923	0.433	-19.007	5.359
	2.00	1.00	-15.975	5.185	0.010	-28.769	-3.182
		3.00	-22.800	4.095	0.000	-32.887	-12.712
	3.00	1.00	6.824	4.923	0.433	-5.3590	19.007
		2.00	22.800	4.095	0.000	12.7128	32.887

1: Bayesian global model after opt.; **2**: Bayesian global model before opt.; **3**: Individual models

Table 4.15- Multiple comparisons (Tamhane's T2 method).

The comparisons between the different classifiers (Table 4.15) confirmed the results obtained with the t-test when applied to the assessment of sensitivity and specificity values.

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
Acc	1.00	2.00	1.720	0.768	0.070	-0.112	3.553
		3.00	3.268	0.768	0.000	1.435	5.102
	2.00	1.00	-1.720	0.768	0.070	-3.553	0.112
		3.00	1.548	0.768	0.115	-0.284	3.381
	3.00	1.00	-3.268	0.768	0.000	-5.102	-1.435
		2.00	-1.548	0.768	0.115	-3.381	0.284

1: Bayesian global model after opt.; **2**: Bayesian global model before opt.; **3**: Individual models

Table 4.16- Multiple comparisons (Tukey method).

Table 4.16 contains the multiple comparisons in relation to accuracy values. The results partially confirmed the results obtained with the t-test. The difference relies on the equality of means between the Bayesian global model after optimization and the Bayesian global model before optimization that should not be rejected. However the significance level (p-*value*=0.07) was very close to the boundary level that considers strong evidence against the null hypothesis (p-*value*=0.05).

Taking into consideration the overall results, it is possible to affirm that the combination approach followed by the optimization step may have potential to improve the risk prediction.

4.2.4 Missing Information

The ability to deal with missing risk factors is one of the major aims of the proposed combination approach.

Table 4.17 contains some test cases to compare three distinct situations: *i*) performance of the model that resulted from the combination scheme (global model) when missing risk factors were replaced by the respective mean values; *ii*) performance of the global model when there was no replacement of missing risk factors; *iii*) performance of complete Cox model (model that contains all variables and all patients) when missing risk factors were replaced by appropriate mean values.

The replacement of missing risk factors was done according to the variables' type, such as:

- Binary variables were replaced successively by values 0 and 1;
- A single imputation method based on the mean value was applied to the variable *age* that is continuous.

It is important to emphasize, that age and sex are variables that are always available in the daily clinical practice. However, in this case age was very useful to test the behaviour of the model when it had to deal with missing risk factors since it was the only continuous variable available in the dataset possible to be considered in this study.

Here, the analysis was focused on the accuracy value[107]. The Bayesian model with no replacement of values presented higher accuracy $(88.8\%;\ CI : 88.6\%\ to\ 88.9\%)$ than the remaining models: Bayesian with replacement $(88.6\%;\ CI : 88.4\%\ to\ 88.7\%)$ and Cox model with replacement $(86.0\%;\ CI : 84.4\%\ to\ 87.7\%)$.

[107] As the dataset is not significantly imbalanced, accuracy is a reliable indicator of the model's performance.

Missing risk factors	Values to replace	Bayes with replacement	Bayes no replacement	Cox with replacement
age	[67]	86.4	89.5	89.80
age + valvular	[67; 0]	86.4	89.4	90.2
	[67; 1]	86.7		85.2
age + cancer	[67; 0]	86.9	88.8	85.3
	[67; 1]	86.4		89.6
age + diabetes	[67; 0]	86.4	89.2	90.6
	[67; 1]	86.5		88.4
age + renal	[67; 0]	86.4	89.4	90.2
	[67; 1]	86.7		85.2
age + valvular + diabetes +	[67;0;0]	86.4	89.6	90.1
	[67;0;1]	86.4		89.1
	[67;1;0]	86.7		86.2
	[67;1;1]	86.9		83.8
age + valvular + diabetes + cancer	[67;0;0;0]	86.4	88.4	88.4
	[67;0;0;1]	86.4		87.7
	[67;0;1;0]	86.4		86.2
	[67;0;1;1]	87.2		89.2
	[67;1;0;0]	86.4		80.7
	[67;1;0;1]	87.5		89.1
	[67;1;1;0]	87.5		76.0
	[67;1;1;1]	86.4		88.2
age + valvular + diabetes + cancer + renal	[67;0;0;0;0]	86.4	88.6	88.8
	[67;0;0;0;1]	86.4		85.6
	[67;0;0;1;0]	86.4		86.4
	[67;0;0;1;1]	86.4		88.8
	[67;0;1;0;0]	86.4		87.5
	[67;0;1;0;1]	86.4		82.5
	[67;0;1;1;0]	86.4		88.4
	[67;0;1;1;1]	86.4		89.3
	[67;1;0;0;0]	86.4		85.6
	[67;1;0;0;1]	87.8		72.9
	[67;1;0;1;0]	86.4		88.8
	[67;1;0;1;1]	86.7		86.3
	[67;1;1;0;0]	86.7		83
	[67;1;1;0;1]	86.7		67.8
	[67;1;1;1;0]	86.4		88.8
	[67;1;1;1;1]	86.4		84.9

Table 4.17 - Accuracy assessment in the presence of missing values (% values).

A statistical analysis was carried out to support some conclusions about the ability of those models to deal with missing risk factors.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
Acc	ea	2.23	0.140	23.9	72	0.000	2.17	0.09	1.99	2.35
ByNR *vs* ByWR				23.9	70.2	0.000	2.17	0.09	1.99	2.35
Acc	ea	24.3	0.000	3.35	72	0.001	2.71	0.81	1.09	4.31
ByNR *vs* Cox				3.35	36.5	0.002	2.71	0.81	1.07	4.34

ByNR: Bayesian global model (from combination scheme) with no replacement; **ByWR:** Bayesian global model (from combination scheme) with replacement; **Cox**: Cox global model (from regression) with replacement

Table 4.18 – Statistical analysis (accuracy).

In both Student's t-tests there is strong evidence against the null hypothesis. Thus, the equality of means should be rejected between the Bayesian global model with no replacement and the Bayesian global model with replacement, as well as between the Bayesian global model with no replacement and the Cox global model.

Cox model had a higher accuracy variance than the Bayesian models, so although the Cox model had performed slightly better than the Bayesian models for some missing risk factors, it showed high performance degradation in other test conditions.

Similarly to the procedure previously adopted the Mann-Whitney U test was carried out (Table 4.19) followed by an analysis of variance.

	Acc ByNR *vs* ByWR	Acc ByNR *vs* Cox
Mann-Whitney U	0.000	455.000
Wilcoxon W	703.000	1158.000
Z	-7.587	-2.503
Asymp. Sig. (2-tailed)	0.000	0.012

Table 4.19- Classifiers comparison (Mann- Whitney U test).

In both cases the null should be rejected which verifies the conclusions obtained with parametric tests.

Table 4.20 assesses the homogeneity of the accuracy variance of the different classifiers.

	Levene Statistic	df1	df2	Sig.
Acc	24.913	2	108	0.000

Table 4.20 - Test of homogeneity of variances.

In this case the homogeneity of variances can not be assumed. This has a direct influence in the analysis of variance procedure.

		Sum of Squares	df	Mean Square	F	Sig.
Acc	Between Groups	152.222	2	76.111	9.436	0.000
	Within Groups	871.115	108	8.066		
	Total	1023.337	110			

Table 4.21 – Comparison of classifiers [ANOVA].

Table 4.21 shows that the equality of means between the several classifiers should be rejected.

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
Acc	1.00	2.00	2.178	0.090	0.000	1.955	2.400
		3.00	2.705	0.806	0.006	0.687	4.723
	2.00	1.00	-2.178	0.090	0.000	-2.400	-1.955
		3.00	0.527	0.805	0.887	-1.489	2.543
	3.00	1.00	-2.705	0.806	0.006	-4.723	-0.687
		2.00	-0.527	0.805	0.887	-2.543	1.489

1: Bayesian global model after opt.; **2**: Bayesian global model before opt.; **3**: Individual models

Table 4.22- Multiple comparisons (Tamhane's T2 Method).

The comparisons between classifiers depicted in Table 4.22 confirmed the previous results.

It is possible to conclude that the Bayesian global model with no replacement had a better performance than the other two classifiers which confirms its superior ability to deal with missing risk factors.

4.3 Tools Applied in Clinical Practice

This validation scenario was developed based on the proper selection and combination of current risk assessment tools. The models obtained through the combination scheme were validated assuming real patient testing datasets made available by two Portuguese hospitals.

4.3.1 Selection of Individual Risk Assessment Tools

Some of the risk assessment tools identified in Table 2.3 were selected for this validation procedure.

Model	Patients Enrolled	Event	Term (months)	Patient's condition	Risk Factors
GRACE *(Tang, 2007)*	1143	Death/MI	6	CAD	Age, SBP, CAA HR, CR, STD, ECE, KIL
PURSUIT *(Boersma, 2000)*	337	Death	1	CAD	Age, Sex, SBP, CCS, HR, STD, ERL, HF
TIMI NSTEMI *(Antman, 2000)*	3171	Death/MI/ UR	14 days	CAD	Age, STD, ECE, KCAD, ASP, ANG, RF

SBP – Systolic blood pressure**, CR**-Creatinine, **HR** – Heart rate, **CAA** – Cardiac arrest at admission, **KIL** – Killip class: II-IV, **STD** - ST segment depression, **ECE** - Elevated cardiac enzymes, **KCAD**- Known coronary artery disease, **ERL** – Enrolment(MI/UA), **HF** –Heart Failure, **CCS** – Angina classification, **ASP** - Use of aspirin in the previous 7 days, **ANG** - 2 or more angina events in past 24 hrs., **RF** - 3 or more cardiac risk factors.

Table 4.23 - Selected risk assessment tools.

All the selected tools were developed to predict events in NSTEMI (non-ST segment elevation myocardial infarction[108]) patients.

[108] Myocardial infarctions (heart attacks) occur when a coronary artery suddenly becomes occluded by a blood clot, causing death to a part of the heart muscle being supplied by that artery. According to their severity myocardial infarctions are divided into two types: STEMI/NSTEMI. A NSTEMI is the less severe type (ST segment elevation indicates that a relatively large amount of heart muscle damage is occurring, because the coronary artery is totally blocked). A more detailed definition of NSTEMI can be found in *(Alpert, 2000)*

4.3.2 Training and Testing Datasets

Leiria-Pombal Hospital Centre (Portugal) and Santa Cruz Hospital (Lisbon/Portugal) provided the real patient datasets that were used as testing datasets in this validation procedure.

Santa Cruz Hospital Testing Dataset

This dataset contains data from N=460 consecutive patients that were admitted in the Santa Cruz Hospital with Acute Coronary Syndrome with non-ST segment elevation (ACS-NSTEMI) from March 1999 to July 2001.

Model	Event
Age (years)	63.4 ± 10.8
Sex (Male/Female)	361 (78.5%) / 99 (21.5%)
Risk Factors:	
Diabetes (0/1)	352 (76.5%) / 108 (23.5%)
Hypercholesterolemia (0/1)	180 (39.1%) / 280 (60.9%)
Hypertension (0/1)	176 (38.3%) / 284 (61.7%)
Smoking (0/1)	362 (78.7 %) / 98 (21.3%)
Previous History / Known CAD:	
Myocardial Infarction (0/1)	249 (54.0%) / 211 (46.0%)
Myocardial Revascularization (0/1)	239 (51.9%) / 221 (48.1%)
PTCA	146 (31.7%)
CABG	103 (22.4%)
Sbp (mmHg)	142.4 ± 26.9
Hr (bpm)	75.3 ± 18.1
Creatinine (mg/dl)	1.37 ± 1.26
Enrolment [0 UA, 1 MI]	180 (39.1 %) / 280 (60.9%)
Killip 1/2/3/4	395 (85.9%) / 31 (6.8%) / 33 (7.3 %) / 0%
CCS [0 I/II; 1 CSS III/IV]	110 (24.0%) / 350 (76.0%)
ST Segment Deviation (0/1)	216 (47.0%) / 244 (53.0%)
Signs of Heart Failure(0/1)	395 (85.9%) / 65 (14.1%)
Tn I > 0.1 ng/ml (0/1)	313 (68.0%) / 147 (32.0%)
Cardiac Arrest Admission (0/1)	460 (100%) / 0%
Aspirin (0/1)	184 (40.0%) / 276 (60.0%)
Angina (0/1)	19 (4.0%) / 441 (96.0%)

Table 4.24 - Risk factors – baseline characteristics (Santa Cruz dataset)[109].

[109] Continuous variables with a normal distribution are expressed as mean value and standard deviation. Discrete variables are presented as frequencies and per cent values

Table 4.25 presents the different endpoints included in the dataset:

Time	Event	N	%	Total
30 days	Death	13	2.8	33
	Myocardial Infarction	24	5.2	7.2%

Table 4.25 - Endpoint rates (Santa Cruz dataset).

Leiria-Pombal Hospital Centre Testing Dataset

The available dataset contains data from N=99 patients[110] that were admitted in the Leiria-Pombal Hospital Centre with Acute Coronary Syndrome with non-ST segment elevation (ACS-NSTEMI) during 2007.

Model	Event
Age (years)	68.0 ± 11.8
Sex (Male/Female)	68 (68.7%) / 31 (31.3%)
Risk Factors:	
Diabetes DMIT (0/1)	91(91.9%) / 8 (8.1%)
Diabetes DMNIT (0/1)	70 (70.7%) / 29 (29.3%)
Hypercholesterolemia (0/1)	59 (59.6%)/ 40 (40.4%)
Hypertension (0/1)	26 (26.3%) / 73 (73.7%)
Smoking (0/1)	83 (83.8) / 16 (16.2%)
Previous History / Known CAD	66 (66.7%) / 33 (33.3%)
Sbp (mmHg)	145.7 ± 32.1
Hr (bpm)	83.2 ± 20.2
Creatinine (mg/dl)	1.11 ± 0.42
Enrolment [0 UA, 1 MI]	6 (6.1%) / 93 (93.9%)
Killip 1/2/3/4	70 (70.7%) / 21 (21.2%) / 7 (7.1%) /1 (1%)
CCS [0 I/II; 1 CSS III/IV]	78 (78.8%) / 21 (21.2%)
ST Segment Deviation (0/1)	98 (99%) / 1 (1%)
Signs of Heart Failure(0/1)	70 (70.7%) / 29 (29.3%)
Tn I > 0.1 ng/ml (0/1)	7 (7.1%) / 92 (92.9%)
Cardiac Arrest Admission (0/1)	98 (99%) / 1 (1%)
Aspirin (0/1)	71 (71.7%) / 28 (28.3%)
Angina (0/1)	33 (33.3%) / 66 (66.7%)

Table 4.26 - Risk factors – baseline characteristics (LPHC dataset).

[110] This is the number of patients with complete information. The patients with missing follow-up information were discarded.

There were 5 events of the observed endpoint (30 days/death), which originated an endpoint rate of 5.1%.

Training Data Set

The derivation of each individual Bayesian classifier that replicates the behavior of each risk assessment tool (Table 4.23) is dependent on the parameter learning procedure (2.44). To achieve this common representation (naïve Bayes classifier), a set of instances $\Upsilon = \{\mathbf{x}_1,...,\mathbf{x}_N\}$ is needed to obtain the required training dataset $D = \{(\mathbf{x}_1, c_1),...,(\mathbf{x}_N, c_N)\}$. Each instance $\mathbf{x}_i = [x_1^i,...,x_p^i]$; $i = 1,...,N$ is applied to each selected risk assessment tool j and the respective c_i^j is obtained. The resulting dataset $D^j = \{(\mathbf{x}_1, c_1^j),...,(\mathbf{x}_N, c_N^j)\}$ is taken to derive the parameters (prior and conditional probabilities) for the j individual Bayesian classifier.

Continuous variables (age, sbp, hr and creatinine) were assumed as normally distributed. Values for the respective mean and standard deviation were taken from literature (Table 4.24; Table 4.26). Some variables were discretized through EWD (age: [30,90] width: 10; creatinine: [0,2.8] width: 0.4) while others were based on clinical significance intervals (sbp: [0,120], [121,140], [141,220]; hr: [0,60], [61,100], [101,220]). Discrete variables are binary and were generated through a random process. The training dataset was created assuming that: $1 \leq i \leq N$; N=1000. The training dataset does not have instances with missing values.

4.3.3 Global Assessment

Individual Risk Assessment Tools

The assessment and posterior combination of individual tools impose that these tools have the same classification goal, i.e. the different tools must have the same number of output categories. Figure 4.2 presents the performed adjustment:

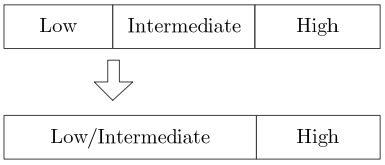


Figure 4.2 - Adjustment of categories (risk assessment tools).

In order to assure the same number of output categories for the three individual models, the original "High" category assumed the value 1 and the remaining categories (low, intermediate, etc.) the value 0. The reduction of output categories was validated by the clinical partner that collaborated in the development of this work:

> *The reduction of output categories (low risk/high risk) is correct. In fact, the aim of cardiologists in clinical practice is to discriminate between high risk patients and low risk patients. In a clinical perspective, the identification of intermediate risk patients is not very significant.*

The performance of the three individual statistical tools identified in Table 4.23 was assessed considering the available testing datasets (Table 4.24, Table 4.26).

Model	%	Santa Cruz 30 days/D/MI	Santa Cruz 30 days/D	LPHC 30 days/D
GRACE	SE	60.6	76.9	60.0
	SP	74.9	73.8	60.6
	Acc	73.9	73.9	60.6
	AUC	0.67	0.765	0.600
PURSUIT	SE	42.4	38.5	20.0
	SP	74.2	73.4	72.3
	Acc	72.0	72.4	69.7
	AUC	0.575	0.565	0.5*
TIMI	SE	33.3	23.1	20.0
	SP	73.5	72.9	93.6
	Acc	70.7	71.5	89.9
	AUC	0.525	0.5*	0.575

SE: Sensitivity; **SP**: Specificity; **Acc**: Accuracy; **AUC:** Area under the ROC curve, **D:** Death; **MI:** Myocardial Infarction; ***** - No discrimination capability.

Table 4.27 – Performance of individual risk assessment tools.

GRACE was the risk assessment tool with the best performance and discrimination capability in the three test situations (Table 4.27). TIMI and PURSUIT presented a poor performance, so they are not as suitable as GRACE to the endpoint prediction in the considered datasets.

Additionally, the assessment of alternative calibrations of individual models through the variation of the respective cut-off values was also carried out.

The first alternative for the recalibration relied on a different combination of original categories as presented in Figure 4.3. Original "intermediate" and "high" categories were grouped in the new "high" category.

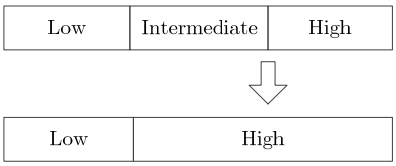


Figure 4.3 - New adjustment of categories.

Table 4.28 presents the obtained results with this approach:

Model	%	Santa Cruz 30 days/D/MI	Santa Cruz 30 days/D	LPHC 30 days/D
GRACE	SE	84.8	84.6	100
	SP	36.5	35.6	38.3
	Acc	40.0	37.0	41.4
	AUC	0.600	0.600	0.650
PURSUIT	SE	90.9	92.3	100.0
	SP	12.2	12.1	8.5
	Acc	17.8	14.3	13.1
	AUC	*0.5**	*0.5**	*0.5**
TIMI	SE	87.9	84.6	100.0
	SP	17.3	17.0	31.9
	Acc	22.4	18.9	35.4
	AUC	*0.525*	*0.520*	*0.650*

SE: Sensitivity; **SP**: Specificity; **Acc**: Accuracy; **AUC:** Area under the ROC curve, **D:** Death; **MI:** Myocardial Infarction; * - No discrimination capability.

Table 4.28 – Performance of individual risk assessment tools (new adjustment).

This option originated a very unbalanced prediction (high sensitivity, very low specificity) which reduced the discrimination capability of these statistical tools. This option was assumed as an unacceptable solution.

The boundaries of categories in the selected risk assessment tools were also evaluated. For instance, the high risk category in TIMI risk score is defined as $\{5,6,7\}$ points *(Antman, 2000)*. Here, the limits were changed, i.e. the high risk category was tested compressed $\{6,7\}$ and stretched $\{4,5,6,7\}$. The same procedure was adopted for the remaining individual tools.

This option did not improve the performance of the individual tools. Hence, the initial approach (Figure 4.2) was adopted: the "high" category of a current risk assessment tool corresponds to the new "high risk", while the remaining categories are grouped in "low risk" category.

Calibration of Bayesian Global Model

The combination strategy was implemented in order to create a global model to perform the risk assessment. Three different testing datasets were considered to evaluate the performance of the global model: *i*) Santa Cruz Hospital, 30 days, combined endpoint: death/myocardial infarction; *ii*) Santa Cruz Hospital, 30 days, endpoint: death; *iii*) Leiria-Pombal Hospital Centre, 30 days, endpoint: death.

1. Weighted Combination – Initial Assessment

Four different test cases were created: *i*) Individual models with the same weight; *ii*) GRACE 100%; PURSUIT 0%; TIMI 0%; *iii*) GRACE 0%; PURSUIT 100%; TIMI 0%; *iv*) GRACE 0%; PURSUIT 0%; TIMI 100%. These testing situations represent the extreme test cases when all the models assume the same weight and when one single model is responsible for all the shared information (prior probabilities/conditional probabilities of variables that belong to more than one model). As expected the weighted combination (*ii*) had better results than the combination that considered the same weight to all individual models (*i*)[111] (Table 4.29). Therefore, the combination scheme should be able to identify the relative importance of the different individual models.

Model	%	Santa Cruz - 30 days/D/MI			
		i	*ii*	*iii*	*iv*
	SE	36.4	*60.6*	*0*	*0*
Global	SP	89.7	67.0	100	100
Model	Acc	85.7	66.5	92.8	92.8
	AUC	0.62	0.64	0.5	0.5

Table 4.29 - Four different testing situations (Santa Cruz dataset, Combined endpoint) [112].

Table 4.30 presents the performance of the global model when intermediate weight combinations were applied to the individual models[113].

[111] The accuracy value in (*i*) is higher than (*ii*) due to the low rate of events (very imbalanced dataset). In imbalanced datasets accuracy is not a reliable metric.

[112] These results were obtained with Santa Cruz Hospital's dataset; combined endpoint. Identical results were achieved with the other datasets as GRACE was the tool with the best performance in all testing datasets.

[113] Based on Table 4.29, the highest weight was assigned to GRACE model.

Model	%	Santa Cruz - 30 days/D/MI				
		i	*ii*	*iii*	*iv*	*v*
	SE	*60.6*	*48.4*	*45.4*	*39.3*	*39.3*
Global	SP	67.0	72.5	77.2	81.9	85.2
Model	Acc	66.5	70.8	75.0	78.9	81.9
	AUC	0.64	0.6	0.55	0.53	0.52

i) GRACE 100%; PURSUIT 0%; TIMI 0%; *ii*) GRACE 90%; PURSUIT 5%; TIMI 5%; *iii*) GRACE 80%; PURSUIT 10%; TIMI 10% *iv*) GRACE 70%; PURSUIT 20%; TIMI 10%; *v*) GRACE 60%; PURSUIT 20%; TIMI 20%;

Table 4.30 - Tested weights of GRACE model.

The testing situation *i* (Table 4.30) registered the best discrimination ability as well as the highest sensitivity value. Therefore, it was adopted through the validation process.

2. Weighted Combination – Bootstrapping Validation

Bootstrapping validation was implemented in order to improve the reliability of the validation results. Bootstrapping validation is based on the statement that if repeated samples are taken from the original sample, simulating the way the data are sampled from the population, then these samples can be used to derive standard errors and confidence intervals of a given parameter. Therefore, this kind of validation allowed the derivation of the confidence intervals for the assessed metrics (sensitivity, specificity).

Dataset	%	Santa Cruz 30 days/D/MI	Santa Cruz 30 days/D	LPHC 30 days/D
Original	*SE*	60.6	61.5	80.0
	SP	67.0	65.7	67.0
	AUC	0.635	0.625	0.725
Bootstrap Samples $N_B = 1000$	*SE*	60.6	61.6	80.3
		CI (60.1;61.3)	*CI (60.7;62.5)*	*CI (78.9;81.5)*
	SP	67.0	65.8	66.8
		CI (66.9;67.2)	*CI (65.6;65.9)*	*CI (66.4;67.2)*

SE: Sensitivity; **SP**: Specificity; **D:** Death; **MI:** Myocardial infarction; *(-;-) = 95% Confidence interval;*

Table 4.31 – Bayesian global model - original samples vs. bootstrap samples.

The results presented in Table 4.31 conclude that validation based on the original dataset and bootstrapping samples had similar results.

Bootstrapping has been implemented to the validation procedure since it is an efficient method of resampling that allows the achievement of more reliable results.

The available testing datasets (SantaCruz hospital and LPHC) are severely imbalanced[114] (low event rates). In this situation, accuracy is not a sensible indicator, for this reason geometric mean was adopted (2.83).

The low event rate of the testing datasets (severely imbalanced) also imposed a restriction regarding other validation strategies, such as cross validation. In fact, the reduced number of events was an additional difficulty that obstructed the random separation of testing instances. This restriction reinforced the selection of bootstrapping validation as the main strategy for the complete validation of the proposed combination methodology.

Bayesian Global Model *vs.* Voting

As previously introduced, the weighted combination strategy was compared with a voting scheme.

Dataset	%	Santa Cruz 30 days/D/MI		Santa Cruz 30 days/D		LPHC 30 days/D	
		ByG	Vot	ByG	Vot	ByG	Vot
Original	SE	60.6	48.5	61.5	53.8	80.0	40.0
	SP	67.0	75.6	65.7	74.7	67.0	74.5
	Gmean	63.4	60.6	63.5	63.0	73.2	54.5
	AUC	0.635	0.625	0.625	0.625	0.725	0.575
Bootstrap Samples $N_B = 1000$	SE	60.6 *(60.1;61.3)*	48.6 *(48.0;49.2)*	61.6 *(60.7;62.5)*	53.7 *(52.9;54.7)*	80.3 *(78.9;81.5)*	41.4 *(40.0;43.1)*
	SP	67.0 *(66.9;67.2)*	75.6 *(75.5;75.8)*	65.8 *(65.6;65.9)*	74.6 *(74.5;74.8)*	66.8 *(66.4;67.2)*	74.1 *(73.7;74.5)*
	Gmean	63.6 *(63.3;63.9)*	60.3 *(60.0;60.7)*	63.1 *(62.7;63.6)*	62.7 *(62.2;63.3)*	72.3 *(71.5;73.1)*	50.6 *(49.3;52.1)*

ByG – Bayesian global model; **Vot** – Voting model

Table 4.32 - Bayesian global model/voting model.

The voting approach to combine model outputs was selected since it can be easily implemented by the physician in the regular clinical practice.

Based on results presented in Table 4.32, it is possible to state that:

[114] As defined by Yen *(Yen, 2009)*, in an imbalanced dataset the *majority class* has a large percentage of all the samples, while the samples in *minority class* just occupy a small part of all the samples. So, a classifier tends to have more ability to predict the *majority class* while it ignores the *minority class*.

- Validations based on original datasets and bootstrap samples present similar results;
- Bayesian global model presents higher sensitivity than the voting model. This is true in the three testing datasets;
- Specificity is higher in the voting model than in the Bayesian global model;
- The discrimination capability of Bayesian global model is higher than the voting model in the LPHC dataset. In regards to Santa Cruz dataset the AUC is similar between the two approaches.

In order to have an additional insight on the performance of both strategies two hypothesis tests were implemented (Student's t-test, Levene's-test)[115]. Parametric tests were adopted regarding the sample size ($N = 1000$) and the statistical independence between data.

These two features allowed the application of the Central Limit theorem which states that the normal distribution provides a good approximation to the sampling distribution of the parameter of interest (sample mean), whatever its underlying distribution, provided that the samples are independent and the sample size is sufficiently large ($N \geq 30$) *(Kirkwood, 2003).*

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	0.003	0.956	30.7	1998	0.000	12.1	0.39	11.31	12.86
ByG *vs* Vot				30.7	1997	0.000	12.1	0.39	11.31	12.86
SP	ea	10.115	0.001	-89.1	1998	0.000	-8,6	0.096	-8,80	-8.42
ByG *vs* Vot				-89.1	1970	0.000	-8.6	0.096	-8.80	-8.42
Gmean	ea	20.052	0.000	13.7	1998	0.000	3.2	0.236	2.77	3.69
ByG *vs* Vot				13.7	1952	0.000	3.2	0.236	2.77	3.69

Table 4.33- Bayesian vs. voting [Santa Cruz dataset (death / myocardial infarction)].

Based on Table 4.33 it is possible to conclude that:

- There is no strong evidence against the equality of the sensitivity's variances. However the equality of means should be rejected, showing that Bayesian approach had higher sensitivity than the voting model;

[115] In both tests, if the *p-value* (significance value) is lower than 0.05, then there is strong evidence against the null hypothesis (null hypothesis should be rejected).

- The homogeneity of variances cannot be assumed for specificity values. The equality of means should also be rejected. In this case the voting model assured the highest specificity value;
- In relation to the geometric mean, homogeneity of variances cannot be assumed. The Bayesian global model had the highest geometric mean value;

Table 4.34 contains the comparisons between the Bayesian global model and voting model considering the Santa Cruz dataset with a different endpoint (death):

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	4.567	0.987	12.1	1998	0.000	7.8447	0.648	6.574	9.115
ByG *vs* Vot				12.1	1998	0.000	7.8447	0.648	6.574	9.115
SP	ea	9.931	0.002	-93.2	1998	0.000	-8.889	0.095	-9.076	-8.702
ByG *vs* Vot				-93.2	1975	0.000	-8.889	0.095	-9.076	-8.702
Gmean	ea	12.19	0.000	1.18	1998	0.236	0.455	0.384	-0.298	1.208
ByG *vs* Vot				1.18	1966	0.236	0.455	0.384	-0.298	1.208

Table 4.34 - Bayesian vs. voting [Santa Cruz dataset (endpoint: death)].

- The sensitivity's homogeneity of variances must be rejected. The Bayesian global model had higher sensitivity than the voting model;
- The voting model had the highest specificity mean value. Similarly to the previous situation, homogeneity of variances must be rejected;
- In relation to geometric mean, the null hypothesis should not be rejected.

The results obtained with the LPHC dataset (Table 4.35) are similar to those extracted from Table 4.34.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	54.886	0.000	37.7	1998	0.000	38.83	1.031	36.81	40.86
ByG *vs* Vot				37.7	1897	0.000	38.83	1.031	36.81	40.86
SP	ea	1.992	0.158	-25.2	1998	0.000	-7.34	0.291	-7.91	-6.78
ByG *vs* Vot				-25.2	1997	0.000	-7.34	0.291	-7.91	-6.78
Gmean	ea	237.1	0.000	26.5	1998	0.000	21.72	0.819	20.11	23.32
ByG *vs* Vot				26.5	1531	0.000	21.72	0.819	20.11	23.32

Table 4.35 - Bayesian vs. voting [LPHC dataset (endpoint: death)].

Although, in this situation the Student's t-test provides strong evidence against the equality of means of the geometric mean. The value is higher in the Bayesian global model than in the voting model.

1. Clinical Usefulness

Clinical usefulness is an important concept that must be addressed to determine which one of the models has more potential to be applied in the regular clinical practice. According to Steyerberg *(Steyerberg, 2009)*, clinical usefulness can be defined as the model's ability to make such classifications better than a default policy without the prediction model. The same author stated that: *missing a patient with the expected outcome is often more important than an incorrect classification of a patient without the outcome*. Thus, in a clinical context false negative errors are usually more important than false positive errors.

The increase of sensitivity[116] is usually more critical than the increase of specificity. In this context, it is possible to affirm that, despite the reduction of specificity, the Bayesian global model had better performance than the voting model.

Bayesian Global Model *vs.* Individual Tools and Voting

Global Bayesian model's performance must also be compared with the performance of individual risk assessment tools. The following tables contain the performance values of all assessed models in the three testing cases.

Dataset	%	GRACE	PURSUIT	TIMI	ByG	Vot.
Original	SE	60.6	42.4	33.3	60.6	48.5
	SP	74.9	74.2	73.5	67.0	75.6
	Gmean	67.3	56.0	49.4	63.4	60.6
	AUC	0.675	0.575	0.525	0.635	0.625
Bootstrap Samples $N_B = 1000$	SE	60.8 *(60.2; 61.3)*	42.4 *(41.9;43.1)*	33.5 *(33.0; 34.0)*	60.6 *(60.1;61.3)*	48.6 *(48.0;49.2)*
	SP	74.9 *(74.8; 75.1)*	74.2 *(74.1;74.3)*	73.6 *(73.5; 73.7)*	67.0 *(66.9;67.2)*	75.6 *(75.5;75.8)*
	Gmean	67.3 *(67.0; 67.6)*	55.8 *(55.5;56.2)*	49.3 *(48.9; 49.7)*	63.6 *(63.3;63.9)*	60.3 *(60.0;60.7)*

Table 4.36 – Bayesian global model/voting model / individual tools [Santa Cruz dataset (D/MI)].

[116] $SE = \frac{TP}{TP + FN}$; $SP = \frac{TN}{TN + FP}$; TP: True Positive; TN: True negative; FN: False negative; FP: False Positive

Dataset	%	GRACE	PURSUIT	TIMI	ByG	Vot.
Original	SE	76.9	38.5	23.1	61.5	53.8
	SP	73.8	73.4	72.9	65.7	74.7
	Gmean	75.3	53.1	40.6	63.5	63.0
	AUC	0.765	0.565	0.5	0.625	0.625
Bootstrap Samples $N_B = 1000$	SE	77.3	38.2	23.0	61.6	53.7
		(76.5;78.0)	*(37.4;39.2)*	*(22.3;23.7)*	*(60.7;62.5)*	*(52.9;54.7)*
	SP	73.8	73.3	72.9	65.8	74.6
		(73.6;73.9)	*(73.1;73.4)*	*(72.8;73.1)*	*(65.6;65.9)*	*(74.5;74.8)*
	Gmean	75.2	51.8	38.8	63.1	62.7
		(74.9;75.6)	*(51.1;52.5)*	*(38.0;39.5)*	*(62.7;63.6)*	*(62.2;63.3)*

Table 4.37 - Bayesian global model/voting model / individual tools [Santa Cruz dataset (death)].

Dataset	%	GRACE	PURSUIT	TIMI	ByG	Vot.
Original	SE	60.0	20.0	20.0	80.0	40.0
	SP	60.6	72.3	93.6	67.0	74.5
	Gmean	60.2	38.0	43.2	73.2	54.5
	AUC	0.6	0.5	0.575	0.725	0.575
Bootstrap Samples $N_B = 1000$	SE	61.2	19.9	21.5	80.3	41.4
		(59.8;62.8)	*(18.6;21.2)*	*(20.3;22.9)*	*(78.9;81.5)*	*(40.0;43.1)*
	SP	60.4	72.1	93.2	66.8	74.1
		(59.9;60.8)	*(71.6;72.5)*	*(92.7;93.5)*	*(66.4;67.2)*	*(73.7;74.5)*
	Gmean	58.7	29.0	35.2	72.3	50.6
		(57.7;59.7)	*(27.4;30.5)*	*(33.4;36.9)*	*(71.5;73.1)*	*(49.3;52.1)*

Table 4.38 - Bayesian global model/voting model / individual tools [LPHC (death)].

PURSUIT and TIMI tools presented a very poor performance namely a very low sensitivity. It is possible to conclude that the Bayesian global model had better behaviour than these individual risk assessment tools. GRACE was the individual tool that presented a competitive performance with the Bayesian global model.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	1.657	0.198	-0.32	1998	0.750	-0.123	0.385	-0.874	0.633
ByG *vs* GRACE				-0.32	1992	0.750	-0.123	0.385	-0.874	0.633
SP	ea	6.664	0.010	-80.7	1998	0.000	-7.894	-8.086	-8.087	-7.703
ByG *vs* GRACE				-80.7	1981	0.000	-7.894	-8.086	-8.087	-7.703
Gmean	ea	0.059	0.808	-17.3	1998	0.000	-3.728	-4.151	-4.151	-3.305
ByG *vs* GRACE				-17.3	1997	0.000	-3.728	-4.151	-4.151	-3.305

Table 4.39 - Bayesian vs. GRACE [Santa Cruz dataset (endpoint: death/myocardial infarction)].

Table 4.39 contains the test data about the comparison between the Bayesian global model and the individual GRACE risk assessment tool. It is possible to conclude that with this dataset GRACE tool had a slightly better behavior than the Bayesian global model, namely:

- In what concerns sensitivity, the homogeneity of variances and equality of mean values may be assumed (*p-value* = 0.198; *p-value* = 0.750), which means that both models achieved a similar sensitivity;
- GRACE tool presented a higher specificity mean value than the Bayesian global model (*Mean Difference* = -7.89). There is also strong evidence against the equality of variances (*p-value* = 0.001).

As a result of the higher specificity, geometric mean was also higher in the GRACE tool than in the Bayesian global model.

The testing dataset (Santa Cruz hospital, endpoint death) registered the highest difference between the performance of the GRACE tool and the Bayesian model.

GRACE had a higher sensitivity (*p-value* = 0.000; *Mean Difference* = -15.69) than the Bayesian global model as well as a higher specificity value (*p-value* = 0.000; *Mean Difference* = -8.01). Accordingly, the geometric mean was also higher in GRACE tool.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	22.617	0.000	-26.2	1998	0.000	-15.69	0.598	-16.86	-14.52
ByG *vs* GRACE				-26.2	1941	0.000	-15.69	0.598	-16.86	-14.52
SP	ea	5.333	0.021	-83.1	1998	0.000	-8.01	0.096	-8.20	-7.822
ByG *vs* GRACE				-83.1	1984	0.000	-8.01	0.096	-8.20	-7.822
Gmean	ea	37.514	0.000	-37.7	1998	0.000	-12.11	0.321	-12.74	-11.48
ByG *vs* GRACE				-37.7	1879	0.000	-12.11	0.321	-12.74	-11.48

ea –equal variances assumed

Table 4.40 - Bayesian vs. GRACE [Santa Cruz dataset (endpoint: death)].

In relation to the dataset of LPHC, the Bayesian global model had a better performance than the GRACE tool:

- The sensitivity's mean value was significantly higher in the Bayesian global model (*p-value* = 0.000; *Mean Difference* = 19.07) than in the GRACE tool. The homogeneity of variance could not be assumed (*p-value* = 0.000);

- The specificity's value was also higher in the Bayesian global model than in the GRACE tool (*p-value* = 0.000; *Mean Difference* = 6.435). There was no strong evidence against the homogeneity of variances;
- Likewise the geometric mean was higher in the Bayesian global model than in the GRACE tool.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	46.579	0.000	18.8	1998	0.000	19.07	1.014	17.089	21.067
ByG *vs* GRACE				18.8	1917	0.000	19.07	1.014	17.089	21.067
SP	ea	0.672	0.412	22.3	1998	0.000	6.435	0.289	5.868	7.001
ByG *vs* GRACE				22.3	1997	0.000	6.435	0.289	5.868	7.001
Gmean	ea	50.337	0.000	21.1	1998	0.000	13.674	0.647	12.406	14.943
ByG *vs* GRACE				21.1	1850	0.000	13.674	0.647	12.406	14.943

Table 4.41 - Bayesian vs. GRACE [LPHC dataset (endpoint: death)].

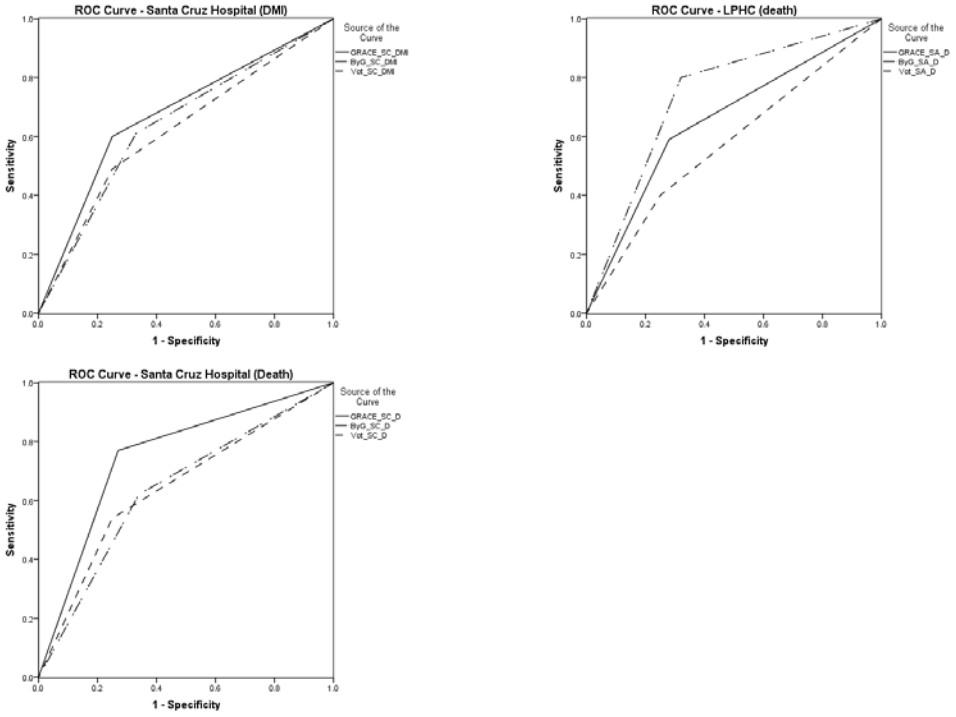


Figure 4.4 – Discrimination capability (area under the ROC curve).

Figure 4.4 presents the ROC curves in the three testing situations. According to the data of Table 4.36; Table 4.37 and Table 4.38, the GRACE tool showed higher discrimination capability when applied to patients of Santa Cruz dataset (combined endpoint; death) while the Bayesian model registered the highest AUC with Leiria-Pombal Hospital Centre's patients.

These results demonstrate that the proposed combination scheme should be complemented with the adjustment of its parameters (optimization procedure) in order to improve its performance.

Bayesian Global Model Before *vs.* After Optimization

An optimization procedure based on genetic algorithm was performed in order to enhance the risk prediction of the global Bayesian model.

Dataset	%	Santa Cruz 30 days/D/MI		Santa Cruz 30 days/D		LPHC 30 days/D	
		ByG	ByG AO	ByG	ByG AO	ByG	ByG AO
Original	SE	60.6	72.7	61.5	76.9	80.0	80.0
	SP	67.0	69.1	65.7	70.7	67.0	82.9
	Gmean	63.4	70.9	63.5	73.7	73.2	81.5
	AUC	0.635	0.7	0.625	0.725	0.725	0.8
Bootstrap Samples $N_B = 1000$	SE	60.6	72.9	61.6	77.3	80.3	79.8
		(60.1;61.3)	*(72.4; 73.4)*	*(60.7;62.5)*	*(76.5; 78.0)*	*(78.9;81.5)*	*(78.6; 81.0)*
	SP	67.0	69.1	65.8	70.6	66.8	83.8
		(66.9;67.2)	*(69.0; 69.2)*	*(65.6;65.9)*	*(70.5 70.8)*	*(66.4;67.2)*	*(83.3; 84.2)*
	Gmean	63.6	70.9	63.1	73.6	72.3	80.9
		(63.3;63.9)	*(70.6; 71.1)*	*(62.7;63.6)*	*(73.3; 74.0)*	*(71.5;73.1)*	*(80.0; 81.6)*

SE: Sensitivity; **SP**: Specificity; **D:** Death; **MI:** Myocardial infarction; ***(-;-)=95% Confidence interval;*** **ByG** – Bayesian global model; **ByG AO** – Bayesian global model after optimization

Table 4.42 - Bayesian global model vs. Bayesian global model after optimization.

Considering the results obtained in Table 4.42, it is possible to conclude that optimization improved the performance of the Bayesian Global model. Statistical significance tests were carried out to support this evidence.

The Student's t-test (Table 4.43) indicated that there was strong evidence against the null hypothesis (equality of means) for the three tested parameters (specificity, sensitivity and geometric mean). These results mean that the optimization algorithm increased the capability of the global model to predict the risk.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	13.602	0.000	-33.0	1998	0.000	-12.27	0.372	-12.99	-11.53
ByG *vs* ByGAO				-33.0	1964	0.000	-12.27	0.372	-12.99	-11.53
SP	ea	1.108	0.293	-20.5	1998	0.000	-2.06	0.101	-2.25	-1.86
ByG *vs* ByGAO				-20.5	1994	0.000	-2.06	0.101	-2.25	-1.86
Gmean	ea	33.731	0.000	-36.7	1998	0.000	-7.29	0.199	-7.68	-6.89
ByG *vs* ByGAO				-36.7	1924	0.000	-7.29	0.199	-7.68	-6.89

Table 4.43 - Bayesian vs. Bayesian after optimization [Santa Cruz dataset (endpoint: D/MI)].

The results presented in (Table 4.44) are similar to Table 4.43.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	22.617	0.000	-26.2	1998	0.000	-15.69	0.598	-16.86	-14.52
ByG *vs* ByGAO				-26.2	1942	0.000	-15.69	0.598	-16.86	-14.52
SP	ea	0.327	0.567	-49.1	1998	0.000	-4.88	0.099	-5.08	-4.69
ByG *vs* ByGAO				-49.1	1997	0.000	-4.88	0.099	-5.08	-4.69
Gmean	ea	44.005	0.000	-32.9	1998	0.000	-10.49	0.319	-11.12	-9.87
ByG *vs* ByGAO				-32.9	1863	0.000	-10.49	0.319	-11.12	-9.87

Table 4.44 - Bayesian vs. Bayesian after optimization [Santa Cruz dataset (endpoint: death)]

The low significance value for the three assessed metrics forced the rejection of equality of means.

The values of mean differences $(-15.69(SE); -4.88(SP); -10.49(G_{mean}))$ show that all of the mean values of the optimized Bayesian global model were higher than those obtained before the optimization.

Table 4.45 presents the results obtained with LPHC dataset. It must be emphasized that in the case of sensitivity ($p\text{-}value = 0.641$) the equality of means should not be rejected, i.e. in this case the optimization algorithm did not improve the sensitivity of risk prediction.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	0.000	0.989	0.47	1998	0.641	0.420	0.899	-1.34	2.18
ByG *vs* ByGAO				0.47	1997	0.641	0.420	0.899	-1.34	2.18
SP	ea	10.769	0.001	-58.1	1998	0.000	-16.96	0.292	-17.53	-16.39
ByG *vs* ByGAO				-58.1	1997	0.000	-16.96	0.292	-17.53	-16.39
Gmean	ea	2.443	0.118	-15.0	1998	0.000	-8.52	0.568	-9.64	-7.40
ByG *vs* ByGAO				-15.0	1988	0.000	-8.52	0.568	-9.64	-7.40

Table 4.45 - Bayesian vs. Bayesian after optimization [LPHC dataset (endpoint: death)].

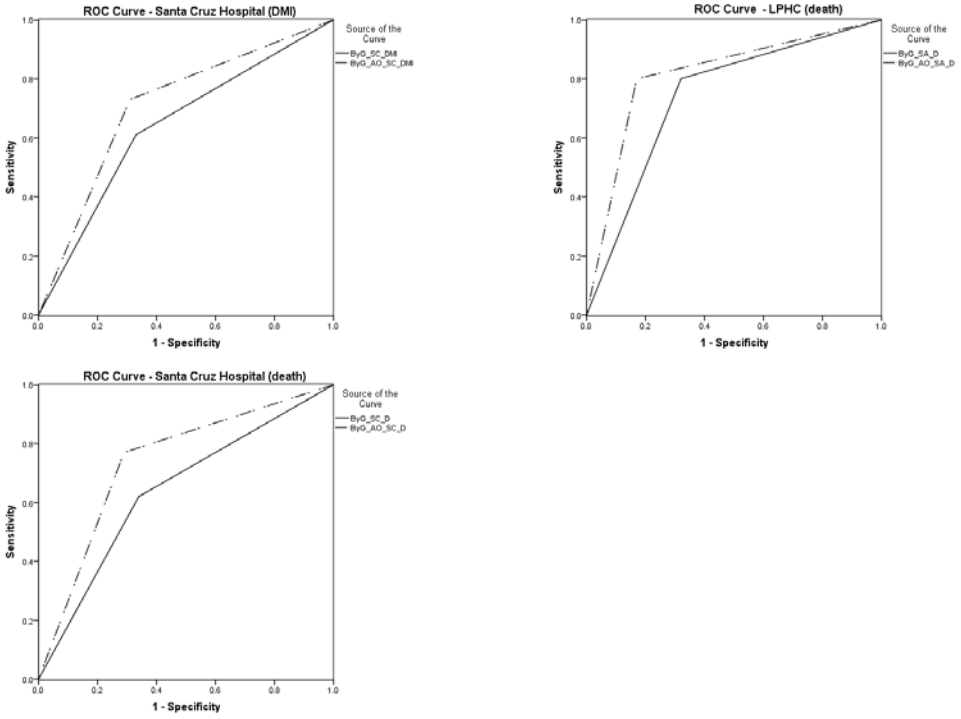


Figure 4.5 - Discrimination capability (Bayesian vs Bayesian after optimization).

Figure 4.5 shows that the optimization procedure increased the discrimination capability of Bayesian global model.

It must be highlighted that the optimization was performed on the neighbourhood of the initial conditional probabilities table's values. After several experiments the value of β_N was defined as $\beta_N = 0.7$. This restriction to the

optimization algorithm operation had two goals: *i*) assure the clinical significance of the model's parameters (conditional probabilities); *ii*) avoid/minimize over-fitting situations.

Finally an analysis of variance was performed to provide a global perspective of the relationships between the several classifiers.

	Levene Statistic	df1	df2	Sig.
SE	6.037	3	3995	0.000
SP	4.200	3	3995	0.006
Gmean	34.761	3	3995	0.000

Table 4.46 - Test of homogeneity of variances [Santa Cruz dataset (endpoint: D/MI)].

As presented in Table 4.46 the homogeneity of variances can not be assumed, which has a direct influence in the analysis of variance procedure.

		Sum of Squares	df	Mean Square	F	Sig.
SE	Between Groups	296603.244	3	98867.748	1385.931	0.000
	Within Groups	284990.088	3995	71.337		
	Total	581593.331	3998			
SP	Between Groups	54567.997	3	18189.332	3929.623	0.000
	Within Groups	18491.947	3995	4.629		
	Total	73059.944	3998			
Gmean	Between Groups	62319.291	3	20773.097	878.905	0.000
	Within Groups	94422.647	3995	23.635		
	Total	156741.938	3998			

Table 4.47 – Comparison of classifiers (ANOVA) [Santa Cruz dataset (D/MI)].

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
SE	1	2	12.267	0.371	0.000	11.288	13.246
		3	24.355	0.370	0.000	23.381	25.330
		4	12.148	0.361	0.000	11.196	13.099
SP	1	2	2.058	0.100	0.000	1.794	2.322
		3	-6.555	0.094	0.000	-6.804	-6.301
		4	-5.836	0.095	0.000	-6.088	-5.584
Gmean	1	2	7.288	0.198	0.000	6.765	7.812
		3	10.521	0.219	0.000	9.944	11.098
		4	3.562	0.197	0.000	3.043	4.081

1: Bayesian global model after opt.; **2**: Bayesian global model before opt.; **3**: Voting model; **4**: GRACE

Table 4.48- Multiple comparisons (Tamhane's T2 method) [Santa Cruz dataset (D/MI)]

Table 4.48 details the comparison between the Bayesian global model after optimization and the remaining classifiers under analysis[117]. It is possible to conclude that the analysis of variance confirms the obtained results through the Student's t test. The same conclusion was obtained with the remaining datasets[118].

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
SE	1	2	15.691	0.598	0.000	14.116	17.266
		3	23.536	0.599	0.000	21.958	25.114
		4	0.000	0.544	1.000	-1.434	1.434
SP	1	2	4.884	0.099	0.000	4.622	5.146
		3	-4.004	0.094	0.000	-4.252	-3.755
		4	-3.127	0.095	0.000	-3.378	-2.875
Gmean	1	2	10.499	0.318	0.000	9.660	11.338
		3	10.954	0.346	0.000	10.040	11.868
		4	-1.614	0.275	0.000	-2.339	-0.890

1: Bayesian global model after opt.; **2**: Bayesian global model before opt.; **3**: Voting model; **4**: GRACE

Table 4.49- Multiple comparisons (Tamhane's T2) [Santa Cruz dataset (endpoint: death)].

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
SE	1	2	-0.420	0.899	0.998	-2.790	1.949
		3	38.417	1.027	0.000	35.712	41.122
		4	18.658	1.010	0.000	15.997	21.318
SP	1	2	16.964	0.292	0.000	16.194	17.733
		3	9.622	0.292	0.000	8.851	10.393
		4	23.399	0.290	0.000	22.633	24.164
Gmean	1	2	8.522	0.567	0.000	7.027	10.018
		3	30.242	0.832	0.000	28.050	32.435
		4	22.197	0.663	0.000	20.449	23.945

1: Bayesian global model after opt.; **2**: Bayesian global model before opt.; **3**: Voting model; **4**: GRACE

Table 4.50- Multiple comparisons (Tamhane's T2) [LPHC dataset (endpoint: death)].

[117] The complete table was not included due to its length.

[118] Only the multiple comparison tables are presented in relation to Santa Cruz (death) and LPHC (death) datasets.

4.3.4 Missing Information

The different classifiers' capability to deal with missing risk factors was assessed through the comparison of the Bayesian approach (before and after the optimization procedure) with the voting model.

Replacement of missing risk factors in the voting model was done according to the variables' type, as follows:

- Binary variables were replaced successively by values 0 and 1;
- Killip level is ordinal it was replaced sequentially by values 1, 2 and 3;
- A single imputation method based on the mean value was applied to the remaining variables that are continuous.

Three different situations were evaluated: *i*) one missing risk factor; *ii*) two missing risk factors; *iii*) three missing risk factors.

Santa Cruz Dataset (endpoint: death / myocardial infarction)

In order to clarify, the assessed metrics are presented in different tables:

Missing Var.	Value	Bayesian *SE %*	Bayesian After Opt. *SE %*	Voting *SE %*	Missing Var.	Value	Bayesian *SE %*	Bayesian After Opt. *SE %*	Voting *SE %*
age	63.4	45.4	48.5	36.3	ccs	0	54.5	63.6	36.3
						1			
sex	0	57.6	54.5	39.3	hfsigns	0	72.2	72.7	45.4
	1			51.5		1			60.6
rf	0	60.6	72.7	48.4	enrol	0	60.6	57.6	30.3
	1			63.6		1			54.5
aspirin	0	54.5	66.7	42.4	killip	1	60.6	75.7	42.4
	1			57.6		2			51.5
kncad	0	54.5	57.6	57.6		3			55.5
	1			54.5	sbp	142.4	54.5	69.9	51.5
angina	0	54.5	60.6	42.4	hr	75.3	*57.6*	66.7	48.5
	1			48.5	creat	1.37	60.6	60.6	48.5
cdarrest	0	60.6	75.7	48.4	stsd	0	54.5	63.6	30.3
	1			54.5		1			51.5
elevated	0	60.6	72.7	42.4					
	1			*66.7*					

Table 4.51 - Sensitivity - one missing risk factor (% values).

Missing Variable	Value	Bayesian	Bayesian After Opt.	Voting
elevated stsd	0; 0	54.5	60.6	27.3
	0; 1			45.4
	1; 0			33.3
	1; 1			69.9

Table 4.52 - Sensitivity - two missing risk factors (% values).

Missing Variable	Value	Bayesian	Bayesian After Opt.	Voting
elevated stsd sbp	0; 0; 142.4	51.5	66.7	27.3
	0; 1; 142.4			45.4
	0; 1; 142.4			30.3
	0; 1; 142.4			72.7

Table 4.53 – Sensitivity - three missing risk factors (% values).

Parameter	Bayesian	Bayesian After Opt.	Voting
Mean	57.1 *CI (55.3;58.8)*	65.4 *CI (62.9;67.7)*	47.8 *CI (43.9;52.6)*
std.	5.1	7.1	11.5
Range	[45.4;72.2]	[48.5;75.7]	[27.3;72.7]

CI= 95% CI

Table 4.54 - Sensitivity - global values (% values).

Two t-tests were performed to compare the Bayesian model after optimization with the remaining models.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	5.70	0.020	5.7	72	0.000	8.230	1.44	5.34	11.1
ByGAO *vs* ByG				5.7	65	0.000	8.235	1.44	5.34	11.1

Table 4.55 – Sensitivity - Bayesian after optimization vs. Bayesian before the optimization.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	5.68	0.020	7.9	72	0.000	17.59	2.23	13.1	22.0
ByGAO *vs* Vot				7.9	72	0.000	17.59	2.23	13.1	22.1

ea – equal variances assumed

Table 4.56 – Sensitivity - Bayesian after optimization vs. voting.

In addition, a one-way ANOVA test was performed to compare the relationships between the sensitivity values originated by the operation of the three different classifiers.

Levene Statistic	df1	df2	Sig
9.850	2	108	0.000

Table 4.57 – Sensitivity - test of homogeneity of variances.

Table 4.57 shows that the condition of homogeneity of variances (*p-value*=0.000) cannot be assumed, although this condition is an ANOVA test's requirement. The procedure to circumvent this situation is identical to the one previously adopted.

Similarly to the results obtained through the Student's t-test, the ANOVA analysis (Table 4.58) showed that the mean values between the three classifiers are different.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5734.83	2	2867.41	40.760	0.000
Within Groups	7597.59	108	70.348		
Total	13332.43	110			

Table 4.58 –Sensitivity - ANOVA analysis.

This information can be detailed through a multiple comparisons test, which assesses the equality of means in all combinations between the three classifiers.

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
SE	1	2	8.235	1.447	0.000	4.689	11.781
		3	17.594	2.231	0.000	12.114	23.074
	2	1	-8.235	1.447	0.000	-11.781	-4.689
		3	9.359	2.082	0.000	4.215	14.503
	3	1	-17.594	2.231	0.000	-23.074	-12.114
		2	-9.359	2.082	0.000	-14.503	-4.215

1 – Bayesian global model after optimization; **2** - Bayesian global model before optimization; **3** – Voting model.

Table 4.59 - Sensitivity - multiple comparisons (Tamhane's T2).

Based on Table 4.59 it is possible to conclude that:

- The equality of means (sensitivity value) should be rejected among all the three classifiers;
- The Bayesian global model after optimization has higher sensitivity than the other two models;
- The Bayesian global model before optimization has higher sensitivity than the voting model.

A similar analysis was performed to the specificity and geometric mean metrics:

Missing Var.	Value	Bayesian	Bayesian After Opt.	Voting	Missing Var.	Value	Bayesian	Bayesian After Opt.	Voting
age	63.4	74.4	75.8	83.4	ccs	0	70.2	68.2	84.1
						1			
sex	0	67.9	75.2	81.0	hfsigns	0	47.3	56.7	73.1
	1			74.7		1			79.1
rf	0	55.2	65.4	65.4	enrol	0	68.8	69.5	85.9
	1			71.4		1			72.1
aspirin	0	70.2	67.9	81.0	killip	1	55.9	48.0	77.7
	1			73.1		2			68.9
kncad	0	69.5	74.7	73.1		3			62.3
	1			72.3	sbp	142.4	67.9	66.75	74.7
angina	0	73.5	72.8	81.5	hr	75.3	67.7	70.5	74.0
	1			75.6	creat	1.37	67.6	68.8	75.6
cdarrest	0	56.2	55.0	75.6	stsd	0	69.3	66.0	89.9
	1			62.0		1			65.6
elevated	0	67.4	69.7	80.3					
	1			64.4					

Table 4.60 - Specificity - one missing risk factor (% values).

Missing Variable	Value	Bayesian	Bayesian After Opt.	Voting
elevated stsd	0; 0	67.9	66.3	93.7
	0; 1			70.7
	1; 0			85.7
	1; 1			46.9

Table 4.61 - Specificity - two missing risk factors (% values).

Missing Variable	Value	Bayesian	Bayesian After Opt.	Voting
elevated stsd sbp	0; 0; 142.4	69.0	66.7	94.8
	0; 1; 142.4			71.2
	0; 1; 142.4			86.6
	0; 1; 142.4			49.6

Table 4.62 - Specificity - three missing risk factors (% values).

Parameter	Bayesian	Bayesian After Opt.	Voting
Mean	65.5 *CI (63.1;67.9)*	65.9 *CI (63.4;68.4)*	74.7 *CI (71.3;78.2)*
std.	7.1	7.4	10.3
Range	[47.3;74.4]	[48.0;75.8]	[46.9; 94.8]

CI= 95% CI

Table 4.63 - Specificity - global values (% values).

The t-test was inconclusive in the comparison between the specificity achieved by the Bayesian classifier before and after the optimization procedure (Table 4.64).

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SP	ea	0.200	0.656	0.3	72	0.789	0.435	1.69	-2.93	3.80
ByGAO *vs* ByG				0.3	71	0.789	0.435	1.69	-2.93	3.80

Table 4.64 – Specificity - Bayesian after optimization vs. Bayesian before optimization.

Table 4.65 shows that there is strong evidence against the equality of means. In this case, the voting model had a higher specificity than the Bayesian model.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SP ByGAO *vs* Vot	ea	2.931	0.091	4.2	72	0.000	-8.80	2.09	-12.9	-4.62
				4.2	65	0.000	-8.80	2.09	-12.9	-4.62

Table 4.65 – Specificity - Bayesian after optimization vs. voting.

Similarly to sensitivity analysis, these comparisons can also be detailed through an analysis of variance. In this case, the assumption of homogeneity of variances is verified (Table 4.66).

Levene Statistic	df1	df2	Sig
2.000	2	108	0.140

Table 4.66 – Specificity - test of homogeneity of variances.

ANOVA analysis shows that mean values between the three classifiers are different:

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2012.919	2	1006.460	14.189	0.000
Within Groups	7660.799	108	70.933		
Total	9673.719	110			

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	Upper Bound
SP	1	2	0.435	1.958	0.973	-4.218	5.088
		3	-8.808	1.958	0.000	-13.461	-4.154
	2	1	-0.435	1.958	0.973	-5.088	4.218
		3	-9.243	1.958	0.000	-13.896	-4.589
	3	1	8.808	1.958	0.000	4.154	13.461
		2	9.243	1.958	0.000	4.589	13.896

Table 4.67 - Specificity - multiple comparisons (Tukey).

The voting model had the highest specificity value in the presence of missing risk factors. Therefore the optimization procedure did not improve the specificity of the

global model in patients with missing risk factors. An identical analysis for geometric mean was performed.

Missing Var.	Value	Bayesian	Bayesian After Opt.	Voting	Missing Var.	Value	Bayesian	Bayesian After Opt.	Voting
age	63.4	58.1	60.7	55.1	ccs	0	61.9	65.9	55.3
						1			
sex	0	62.5	64.0	56.5	hfsigns	0	58.6	64.2	63.1
	1			62.0		1			59.9
rf	0	57.8	69.0	69.0	enrol	0	64.6	63.4	64.6
	1			67.4		1			51.0
aspirin	0	61.9	67.3	58.6	killip	1	58.2	60.3	62.7
	1			64.9		2			57.4
kncad	0	61.5	65.6	64.9		3			59.5
	1			63.6	sbp	142.4	60.9	68.2	62.0
angina	0	63.3	66.5	58.8	hr	75.3	*62.4*	68.6	59.9
	1			60.6	creat	1.37	67.6	68.8	60.6
cdarrest	0	58.3	62.1	60.6	stsd	0	61.5	64.8	52.2
	1			58.2		1			58.1
elevated	0	63.9	71.2	58.4					
	1			*65.5*					

Table 4.68 – Geometric mean - one missing risk factor (% values).

Missing Variable	Value	Bayesian	Bayesian After Opt.	Voting
elevated stsd	0; 0	60.9	63.4	50.5
	0; 1			56.7
	1; 0			53.4
	1; 1			57.2

Table 4.69 - Geometric mean - two missing risk factors (% values).

Missing Variable	Value	Bayesian	Bayesian After Opt.	Voting
elevated stsd sbp	0; 0; 142.4	59.6	65.5	50.9
	0; 1; 142.4			56.8
	0; 1; 142.4			51.2
	0; 1; 142.4			60.0

Table 4.70 – Geometric mean - three missing risk factors (% values).

Parameter	Bayesian	Bayesian After Opt.	Voting
mean	61.0 *CI (60.2;61.8)*	65.2 *CI (64.2;66.0)*	59.0 *CI (57.5;60.6)*
std.	2.3	2.8	4.7
range	[57.8; 67.6]	[60.3; 71.2]	[50.5; 69.0]

Table 4.71 – Geometric mean- global values (% values).

Like the previous situations, t-tests were performed and were complemented with an ANOVA analysis.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
G*mean* ByGAO *vs* ByG	ea	0.949	0.333	6.9	72	0.000	4.15	0.59	2.95	5.34
				6.9	69	0.000	4.15	0.59	2.95	5.34
G*mean* ByGAO *vs* Vot	ea	7.114	0.009	6.8	72	0.000	6.09	0.89	4.30	7.88
				6.8	58	0.000	6.09	0.89	4.30	5.88

ea – equal variances assumed

Table 4.72 – Geometric mean - Bayesian after optimization vs. Bayesian before optimization and Bayesian after optimization vs. voting.

The Bayesian global model registered the highest geometric mean in the presence of missing risk factors. The analysis of variance confirmed this conclusion. Here, the assumption of homogeneity of variances is not verified.

Levene Statistic	df1	df2	Sig
7.895	2	108	0.001

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	716.6	2	358.3	30.56	0.000
Within Groups	1266.0	108	11.7		
Total	1982.7	110			

Table 4.73- Geometric mean - ANOVA analysis.

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
Gmean	1	2	4.151	0.599	0.000	2.685	5.617
		3	6.091	0.897	0.000	3.887	8.296
	2	1	-4.151	0.599	0.000	-5.617	-2.685
		3	1.940	0.858	0.082	-0.176	4.057
	3	1	-6.091	0.897	0.000	-8.296	-3.887
		2	-1.940	0.858	0.082	-4.057	0.1765

1 – Bayesian global model after Optimization; **2** - Bayesian global model before Optimization; **3** – Voting model

Table 4.74 – Geometric mean – Tamhane's T2 analysis.

As expected, the Bayesian global model after optimization was the classifier with the highest geometric mean value.

Santa Cruz Dataset (endpoint: death)

The testing procedure is similar to the one adopted with the combined endpoint. The presentation of all the validation data would be very extensive, for that reason only the global values and ANOVA results are presented.

	Parameter	Bayesian	Bayesian After Opt.	Voting
SE	mean	60.0 *(58.1;61.9)*	63.0 *(61.0;65.1)*	49.6 *(45.2;54.1)*
	std.	5.6	7.0	13.2
	range	[46.1;76.9]	[46.2;84.6]	[23.0;76.9]
SP	mean	64.6 *(62.2,66.9)*	68.8 *(66.8,70.8)*	74.3 *(70.9,77.8)*
	std.	6.9	5.8	10.2
	range	[46.5, 72.9]	[55.0, 78.0]	[46.0, 93.7]
G*mean*	mean	62.0 *(61.0;63.0)*	65.2 *(63.5;66.8)*	59.5 *(57.5;61.6)*
	std.	2.9	4.9	6.1
	range	[57.9;66.8]	[58.7;75.1]	[46.2;70.6]

Table 4.75 - Global values (one, two and three missing risk factors) Santa Cruz dataset [death].

	Levene Statistic	df1	df2	Sig
SE	15.520	2	108	0.000
SP	3.137	2	108	0.047
G*mean*	7.992	2	108	0.001

		Sum of Squares	df	Mean Square	F	Sig.
SE	Between Groups	3653.8	2	1826.9	15.96	0.000
	Within Groups	12356.1	108	114.4		
	Total	16010.0	110			
SP	Between Groups	1764.3	2	882.1	14.049	0.000
	Within Groups	6781.7	108	62.7		
	Total	8546.1	110			
G*mean*	Between Groups	587.9	2	293.9	12.54	0.000
	Within Groups	2530.3	108	23.4		
	Total	3118.2	110			

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
SE	1	2	2.970	2.123	0.424	-2.268	8.209
		3	13.381	2.899	0.000	6.289	20.472
	2	1	-2.970	2.123	0.424	-8.209	2.268
		3	10.410	2.373	0.000	4.541	16.280
	3	1	-13.381	2.899	0.000	-20.472	-6.289
		2	-10.410	2.373	0.000	-16.280	-4.541
SP	1	2	4.167	1.497	0.021	0.503	7.832
		3	-5.564	1.942	0.017	-10.342	-0.787
	2	1	-4.167	1.497	0.021	-7.832	-0.503
		3	-9.732	2.041	0.000	-14.738	-4.726
	3	1	5.564	1.942	0.017	0.787	10.342
		2	9.732	2.041	0.000	4.726	14.738
Gmean	1	2	3.145	0.943	0.004	0.827	5.464
		3	5.624	1.290	0.000	2.467	8.781
	2	1	-3.145	0.943	0.004	-5.464	-0.827
		3	2.478	1.115	0.089	-0.274	5.231
	3	1	-5.624	1.290	0.000	-8.781	-2.467
		2	-2.478	1.115	0.089	-5.231	0.274

1 – Bayesian global model after optimization; **2** - Bayesian global model before optimization; **3** – Voting model
Tamhane's T2 was applied to perform the multiple comparisons test.

Table 4.76 –ANOVA analysis – Santa Cruz dataset (death).

Some conclusions can be derived based on results presented in Table 4.76:

- The Bayesian global model after optimization presented the highest sensitivity value. However, the equality of means should not be rejected in relation to Bayesian global model before optimization. This means that the optimization had a limited effect in the sensitivity value's improvement;
- Similarly to the previous test case (Santa Cruz dataset with combined endpoint), the voting model presented the highest specificity value. ANOVA demonstrated that the optimization procedure slightly increased the prediction specificity value of the Bayesian global model;
- The Bayesian global model after optimization also registered the highest value of geometric mean. In both situations (comparisons with classifiers 2 and 3) the equality of means should be rejected. This reinforces that the optimization procedure increased the geometric mean.

Leiria-Pombal Hospital Centre Dataset (endpoint: death)

The validation approach was similar to the one applied in the Santa Cruz dataset. The global values are presented as well as the ANOVA results.

	Parameter	Bayesian	Bayesian After Opt.	Voting
SE	mean	70.8 *(66.4;75.1)*	75.1 *(71.1;79.1)*	45.4 *(38.4;52.3)*
	std.	12.9	11.9	20.8
	range	[20;80]	[20;80]	[20;100]
SP	mean	65.5 *(64.4;66.6)*	79.4 *(78.1;80.7)*	73.5 *(70.9;76.1)*
	std.	3.3	3.9	7.8
	range	[61.7;81.9]	[75.5;81.9]	[58.5;88.9]
G*mean*	mean	62.1 *(61.0;63.0)*	65.2 *(63.5;66.8)*	59.5 *(57.5;61.6)*
	std.	2.9	4.9	6.1
	range	[57.9;66.8]	[58.7;75.1]	[46.2;70.6]

Table 4.77 - Global values (one, two, three missing risk factors) LPHC dataset [death].

	Levene Statistic	df1	df2	Sig
SE	4.219	2	108	0.017
SP	10.042	2	108	0.000
G*mean*	3.177	2	108	0.046

Table 4.78 - Test of homogeneity of variances - LPHC dataset (death).

		Sum of Squares	Df	Mean Square	F	Sig.
SE	Between Groups	19091.8	2	9545.9	38.29	0.000
	Within Groups	26918.9	108	249.2		
	Total	46010.8	110			
SP	Between Groups	3606.8	2	1803.4	61.2	0.000
	Within Groups	3179.3	108	29.4		
	Total	6786.2	110			
G*mean*	Between Groups	8157.8	2	4078.9	66.5	0.000
	Within Groups	6619.8	108	61.2		
	Total	14777.6	110			

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% CI Lower Bound	95% CI Upper Bound
SE	1	2	4.324	2.899	0.365	-2.764	11.414
		3	29.729	3.955	0.000	20.000	39.459
	2	1	-4.324	2.899	0.365	-11.414	2.765
		3	25.405	4.045	0.000	15.470	35.339
	3	1	-29.729	3.955	0.000	-39.459	-20.000
		2	-25.405	4.045	0.000	-35.339	-15.470
SP	1	2	-13.910	0.844	0.000	-15.977	-11.844
		3	-8.000	1.402	0.000	-11.470	-4.529
	2	1	13.910	0.844	0.000	11.844	15.977
		3	5.910	1.446	0.000	2.344	9.477
	3	1	8.000	1.402	0.000	4.529	11.470
		2	-5.910	1.446	0.000	-9.477	-2.344
Gmean	1	2	9.148	1.522	0.000	5.427	12.870
		3	20.943	1.958	0.000	16.139	25.747
	2	1	-9.148	1.522	0.000	-12.870	-5.427
		3	11.794	1.946	0.000	7.019	16.571
	3	1	-20.943	1.958	0.000	-25.747	-16.139
		2	-11.794	1.946	0.000	-16.570	-7.019

1 *– Bayesian global model after Optimization;* ***2*** *- Bayesian global model before Optimization;* ***3*** *– Voting model*
Tamhane's T2 was applied to perform the multiple comparisons test.

Table 4.79 –ANOVA analysis – LPHC dataset (death).

In this case the conclusions are similar to the previous datasets. The optimization procedure increased all the assessed metrics.

4.3.5 Software Application

The current risk assessment tools combination was validated with implemented software in Matlab. Figure 4.6 presents a graphical interface that allows the definition of two main types of information: *i*) individual model weights; *ii*) risk factors that integrate the global model.

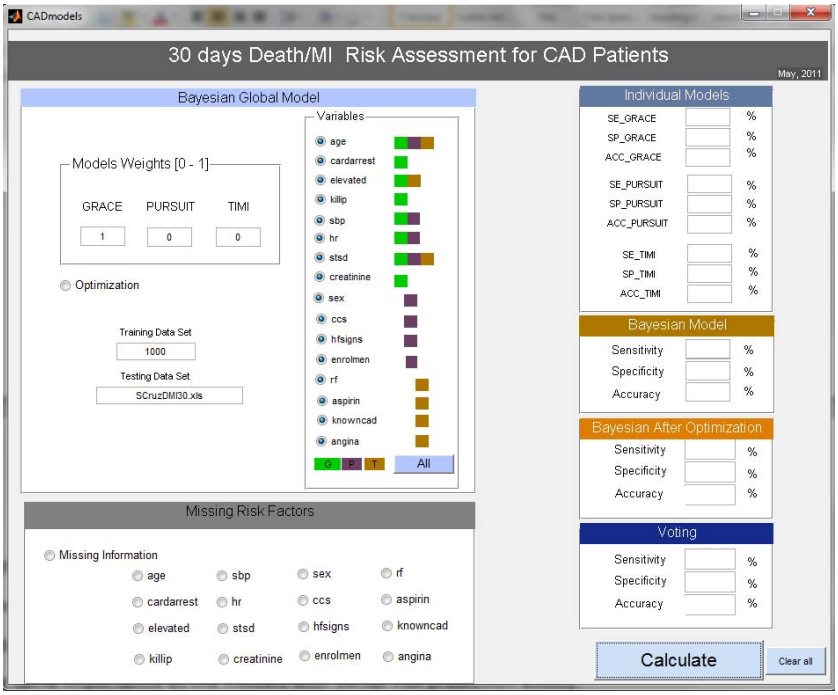


Figure 4.6 – Software to validate the combination methodology.

Therefore, the physician can:

- Define the same weight to all individual risk assessment tools. In this situation there is no distinction among the capability of individual tools to predict the risk of death/MI in that specific population. In this case, the tools with poor performance have the same importance as tools with better risk prediction ability;
- Define different weights based on previous knowledge on individual tools' performance. This method relies directly on the physician's knowledge. This option can potentially achieve better results than the previous one, since it may reduce the contributions of the individual tools with poor behavior;

- The physician also has the capability of choosing the variables that integrate the model. For instance, cardiac arrest may not be considered since there are no cases of cardiac arrest among the potential candidates (patients) for risk assessment. This permits the validation of missing information;
- Finally an optimization procedure can be triggered if the respective option is activated. Otherwise, the parameters that were originated for the last optimization in that specific population are applied.

The software presented in Figure 4.6 was developed to validate the combination methodology, i.e. to allow the comparison between the performances of the several models (Bayesian global model before/after optimization, individual tools, voting model). Software to assess the risk of an individual patient was also implemented as represented in Figure 4.7.

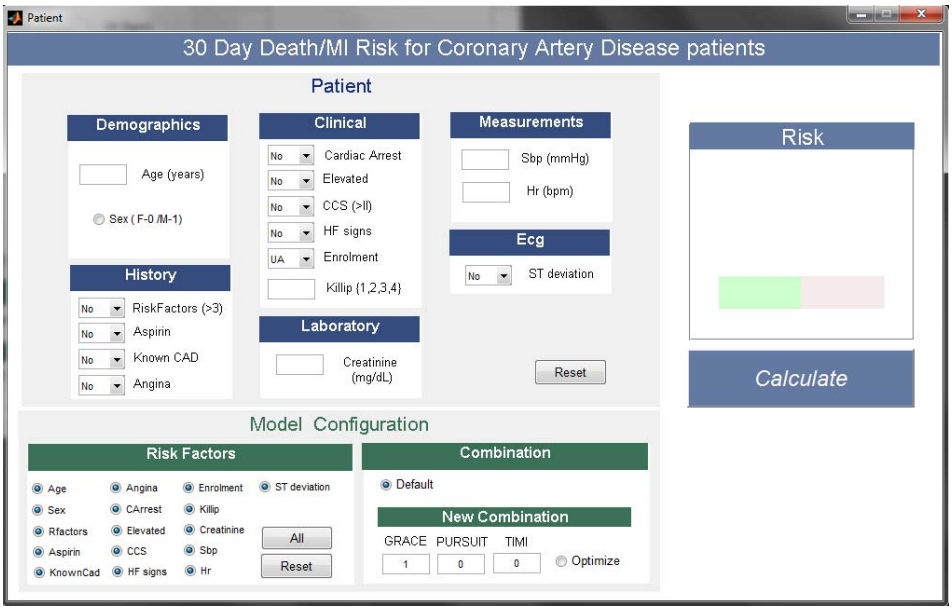


Figure 4.7 – Software to assess the individual patient's risk.

In this case, the physician can:

- Define the individual patient's data. All the variables may be set as missing risk factors. The exception is the information about the patient's sex as this variable is always available in the daily clinical practice;
- Configure the Bayesian global model, i.e. select the risk factors that integrate the model as well as define the conditional probability tables. Actually, the physician may load a previously optimized CPT (default) or define new

weights for the individual risk assessment tools. Optionally, after the new weights definition, an optimization procedure may be triggered to adjust the parameters of the global model.

4.4 Incorporation of Clinical Knowledge

The body mass index (BMI) was selected as the new risk factor to be incorporated in the risk prediction.

The World Health Organization *(WHO , 2011)*, defines BMI as a simple index of weight-for-height that is commonly used to classify underweight, overweight and obesity in adults.It is calculated as follows:

$$BMI = \frac{weight(kg)}{height^2(m)} \tag{4.1}$$

Table 4.80 presents the categories defined based on BMI values:

Classifier	BMI (kg / m^2)	
	Principal cut-off points	Additional cut-off points
Underweight	<18.5	<18.5
Severe thinness	<16.0	<16.0
Moderate thinness	16.0 – 16.99	16.0 – 16.99
Mild thinness	17.0 – 18.49	17.0 – 18.49
Normal Range	18.5 – 24.99	18.5 – 22.99
		23.0 – 24.99
Overweight	≥25.0	≥25.0
Pre-obese	25.0 – 29.99	25.0 – 27.49
		27.5 – 29.99
Obese	≥30.0	≥30.0
Obese class I	30.0 – 34.99	30.0 – 32.49
		32.5 – 34.99
Obese class II	35.0 – 39.99	35.0 – 37.49
		37.5 – 39.99
Obese class III	≥40.0	≥40.0

Table 4.80- The international BMI classification of an adult (WHO, 2011).

A direct health consequence of obesity is a major risk for cardiovascular disease *(WHO , 2011)*. However, some recent studies show that patients who are underweight also have an increased risk of death *(Zheng, 2011)*.

The prevalence in Portugal of the weight categories *(Carmo, 2006)* is presented in Table 4.81:

	Women		Men		Total	
	N	%	N	%	N	%
Low (<18.5)	126	3.4	27	1.0	153	2.4
Normal (18.5-24.9)	1830	49.4	1069	39.5	2899	45.2
Overweight(25.0 – 29.9)	1256	33.9	1216	44.9	2472	38.6
Obesity I (30.0-34.9)	371	10.0	341	12.6	712	11.1
Obesity II (35.0-39.9)	89	2.4	49	1.8	138	2.1
Obesity III (≥40)	33	0.9	5	0.2	38	0.6

Table 4.81 - Prevalence of BMI categories in adults (18-64 years) in 2003-2005 survey.

The BMI's conditional probabilities table (Table 4.82) must reflect not only the BMI prevalence but also the risk associated with each one of the considered categories (underweight, normal/overweight, obese)[119].

	Low Risk	High Risk
BMI ≤ 18.5	0.01	0.02
18.5 < BMI < 30	0.89	0.78
BMI ≥ 30	0.1	0.2
	$\sum_{i=1}^{n_i} P(X_i = x_i^k \mid C = LR) = 1$	$\sum_{i=1}^{n_i} P(X_i = x_i^k \mid C = HR) = 1$

Table 4.82 - BMI's conditional probabilities table[120].

The testing dataset used for this validation was extracted from Santa Cruz Hospital (combined endpoint) (Table 4.24). Thirty-four from the total of 460 patients do not have BMI value (height value was not measured) consequently these patients were not included in this analysis.

The validation methodology presented in Figure 3.7 was implemented, e.g. Table 4.83 shows the conditional probabilities table of GRACE model.

[119] Categories defined by the clinical partner from Santa Cruz Hospital.

[120] This table was defined after several experiments.

Risk factor	LOW	HIGH	Risk factor	LOW	HIGH
age	0.0329	0.0064	hr	0.6667	0.5934
	0.1831	0.0610		0.2864	0.3291
	0.3803	0.2287		0.0469	0.0775
	0.3286	0.3761	stsd	0.7371	0.4562
	0.0657	0.2529		0.2629	0.5438
	0.0094	0.0750	creatinine	0.1033	0.0851
cardarrest	0.8498	0.4003		0.1643	0.1156
	0.1502	0.5997		0.1831	0.1626
elevated	0.6291	0.4587		0.1690	0.1792
	0.3709	0.5413		0.1690	0.1614
killip	0.5211	0.1423		0.2	0.2897
	0.3099	0.2567		0.011	0.0064
	0.1174	0.2821		***0.01***	***0.02***
	0.0516	0.3189	***bmi***	***0.89***	***0.78***
sbp	0.0329	0.0673		***0.1***	***0.2***
	0.2723	0.4396			
	0.6948	0.4930			

Table 4.83 - GRACE + BMI conditional probabilities table.

Table 4.84 contains the results obtained with the three individual Bayesian models before and after the BMI incorporation.

	GRACE	GRACE +BMI	PURSUIT	P.+BMI	TIMI	TIMI+BMI
SE	58.1	61.2	42.1	48.7	26.3	30.6
%	*(57.4; 58.6)*	*(60.6; 61.8)*	*(41.6; 42.7)*	*(48.2; 49.3)*	*(26.1; 26.4)*	*(30.5; 30.8)*
SP	74.8	70.5	74.4	72.6	67.5	57.7
%	*(74.7; 75.0)*	*(70.4; 70.7)*	*(74.2; 74.5)*	*(72.5; 72.8)*	*(67.0; 68.1)*	*(57.2; 58.2)*
Gmean	65.7	65.5	55.7	59.2	42.0	41.9
%	*(65.3; 66.0)*	*(65.2; 65.8)*	*(55.3; 56.0)*	*(58.9; 59.6)*	*(41.8; 42.2)*	*(41.7; 42.1)*

Bootstrap Samples: $N_B = 1000$ *;* (-;-) = 95% Confidence Interval;

Table 4.84 - Results of the BMI's incorporation (Santa Cruz dataset, combined endpoint).

Student's t-tests were performed in order to clarify the influence of the incorporation of the BMI risk factor.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	0.795	0.373	-7.6	1998	0.000	-3.15	0.412	-3.96	-2.34
Gr vs. GrBmi				-7.6	1996	0.000	-3.15	0.412	-3.96	-2.34
SP	ea	0.445	0.505	42	1998	0.000	4.31	0.101	4.11	4.51
Gr vs. GrBmi				42	1996	0.000	4.31	0.101	4.11	4.51
G*mean*	ea	6.08	0.014	0.7	1998	0.445	0.179	0.234	-0.28	0.63
Gr vs. GrBmi				0.7	1984	0.445	0.179	0.234	-0.28	0.63

ea –equal variances assumed; Gr – GRACE; GrBmi – GRACE + BMI

Table 4.85 - GRACE vs. GRACE+BMI.

The BMI incorporation had some influence on the performance of the initial model (GRACE). Sensitivity was improved *(Mean Difference = -3.15)*, however the specificity was reduced *(Mean Difference = 4.31)*. As a result from these two opposite effects the null hypothesis should not be rejected (equality of means) in relation to geometric mean.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	0.003	0.959	-16	1998	0.000	-6.60	0.395	-7.37	-5.82
Ps vs. PsBmi				-16	1998	0.000	-6.60	0.395	-7.37	5.82
SP	ea	0.035	0.852	17	1998	0.000	1.75	0.101	1.55	1.94
Ps vs. PsBmi				17	1996	0.000	1.75	0.101	1.55	1.94
G*mean*	ea	7.190	0.007	-14	1998	0.000	-3.57	0.255	-4.07	-3.07
Ps vs. PsBmi				-14	1983	0.000	-3.57	0.255	-4.07	-3.07

ea –equal variances assumed; Ps – PURSUIT; PsBmi – PURSUIT + BMI

Table 4.86 - PURSUIT vs. PURSUIT+BMI.

The incorporation of the new risk factor improved the sensitivity *(Mean Difference = -6.60)* as well as the geometric mean *(Mean Difference = -3.75)* although it slightly reduced the specificity *(Mean Difference = 1.75)* of the PURSUIT model. The equality of means should be rejected in all three metrics.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	0.083	0.773	42	1998	0.000	4.35	0.103	4.15	4.55
Ti *vs* TiBmi				42	1997	0.000	4.35	0.103	4.15	4.55
SP	ea	0.088	0.766	-24	1998	0.000	-9.84	0.395	-10.6	-9.07
Ti *vs* TiBmi				-24	1998	0.000	-9.84	0.395	-10.6	-9.07
G*mean*	ea	4.935	0.026	-0.7	1998	0.470	-0.11	0.158	-0.42	0.19
Ti *vs* TiBmi				-0.7	1998	0.470	-0.11	0.158	-0.42	0.19

ea –equal variances assumed; Ti – TIMI; TiBmi – TIMI + BMI

Table 4.87 - TIMI vs. TIMI+BMI.

The TIMI model also had a poor performance (Table 4.87). The incorporation of BMI had a positive impact on the sensitivity value. However it was overturned by the decrease of specificity, which was reflected in the geometric mean value.

The validations considering the remaining datasets matched the conclusions obtained with Santa Cruz dataset (combined endpoint).

The obtained results show that the incorporation of additional clinical knowledge may have a positive impact on the risk prediction.

4.5 Personalization based on Grouping of Patients

Two scenarios were explored for the validation of the personalization based on the grouping of patients' methodology: *i*) simulation – theoretical individual models; *ii*) tools applied in clinical practice.

4.5.1 Simulation – Theoretical Individual Models

The first step of this validation scenario was the selection of the theoretical models that were previously derived, through Cox regression, from the TEN-HMS dataset (Table 4.3). The selection process was carried out according to the individual models' accuracy (Table 4.5). After several experiments, three models were selected (*M10*, *M12*, *M22*). It is important to refer, that other models could have been chosen.

However the selected set of models seemed appropriate since it assured some diversity in the classification achieved by the individual models.

A set of $N = 1000$ instances comprising $p = 12$ variables (Table 4.2) was generated according to the procedure depicted in Section 4.2.1.

This dataset $\Upsilon = \{\mathbf{x}_1,...,\mathbf{x}_N\}$ was applied to each one of the individual models, where each model provided a risk probability for each patient. Thus, the dimensionality reduction was implemented based on the three individual model outputs y_i where $0 \le y_i \le 1$. This allowed the mapping between the original data $\mathbf{X}_{12\times1000}$ and the reduced dimensional space $\mathbf{Y}_{3\times1000}$.

The subtractive clustering was applied to $\mathbf{Y}_{3\times1000}$ aiming for the creation of the groups (clusters) of patients. The adjustment of the clustering algorithm parameters (Equations (2.76), (2.77)) was performed through an iterative testing procedure[121]. The main goal of the clustering process (grouping of patients) was to find a trade-off between the number of clusters/respective dimension and the performance (SE/SP) achieved in the classification process[122]. After several experiments $K = 26$ clusters were created, Table 4.88 presents the values of SE/SP obtained by each individual model in each cluster. This number of clusters makes more difficult the clinical interpretability of the grouping of patients. However it does not affect the automatic identification of the cluster that a given patient belongs to.

The *true data* needed to compute these metrics was obtained through the application of $\Upsilon = \{\mathbf{x}_1,...,\mathbf{x}_N\}$ to the complete Cox model. For each patient of each cluster the output class of each model[123] was compared with the *true data.* Thus it was possible to evaluate the performance (SE/SP) of each individual model in each group of patients.

As a result, the several individual models were assigned to the different clusters according to the algorithm detailed in Figure 3.9.

---

[121] The clustering algorithm was implemented based on the Statistics Toolbox – Matlab.

[122] Clusters with low dimension inhibit the application of the main concept (group of patients) of the proposed methodology. The opposite situation (clusters with high dimension) may degradate the performance of the risk assessment tools in those clusters.

[123] As defined in Section 4.2.1., each model's output (risk) had two possible values (low/intermediate risk $\le 30\%$, high risk $> 30\%$). This cut-off value can be easily adjusted.

C	*M10*		*M12*		*M22*		P	E
	SE	SP	SE	SP	SE	SP		
1	21.4	84.2	64.3	71.1	14.3	100.0	104	28
2	100.0	0.0	100.0	0.0	31.3	100.0	44	32
3	100.0	0.0	75.0	0.0	91.7	100.0	50	48
4	100.0	0.0	100.0	0.0	53.6	0.0	56	56
5	20.0	0.0	100.0	0.0	73.3	0.0	30	30
6	100.0	0.0	100.0	0.0	100.0	0.0	32	32
7	100.0	0.0	100.0	0.0	93.1	0.0	58	58
8	100.0	0.0	100.0	0.0	100.0	0.0	54	54
9	100.0	0.0	53.6	0.0	100.0	0.0	56	56
10	100.0	0.0	100.0	0.0	100.0	0.0	30	30
11	100.0	0.0	100.0	0.0	100.0	0.0	38	38
12	100.0	0.0	0.0	100.0	94.1	33.3	46	34
13	100.0	8.3	0.0	100.0	50.0	83.3	28	4
14	100.0	0.0	100.0	0.0	0.0	100.0	28	24
15	100.0	0.0	71.4	0.0	100.0	0.0	28	28
16	76.9	0.0	100.0	0.0	100.0	0.0	26	26
17	100.0	0.0	100.0	0.0	100.0	0.0	26	26
18	100.0	0.0	100.0	0.0	100.0	0.0	42	42
19	100.0	0.0	78.9	0.0	100.0	0.0	38	38
20	14.3	40.0	100.0	0.0	0.0	100.0	24	14
21	100.0	0.0	100.0	14.3	0.0	100.0	16	2
22	10.0	0.0	100.0	0.0	40.0	0.0	20	20
23	56.3	0.0	87.5	100.0	100.0	0.0	34	32
24	100.0	0.0	100.0	0.0	100.0	0.0	38	38
25	100.0	0.0	100.0	0.0	100.0	0.0	40	40
26	0.0	85.7	0.0	100.0	0.0	100.0	14	0

C: cluster; SE: sensitivity (%); SP: specificity (%); P: number of patients; E: number of events

Table 4.88- Performance of selected individual simulated models in each cluster

The derivation of the testing dataset followed the same procedure that was adopted in this first phase. A set of instances $\Upsilon_s = \{\mathbf{x}_1,...,\mathbf{x}_N\}$, $N = 1000$ was generated and it was applied to the complete Cox model in order to obtain the *true data*. Each instance $\mathbf{x}_i \in \Upsilon_s$ was assigned to a specific cluster and it was classified by the individual model that best classifies the patients that belong to that cluster.

The derivation of the testing datasets was repeated $n = 30$ times with the objective of enhancing the statistical significance of the obtained results.

Table 4.89 presents the main results obtained with this validation scenario. It is possible to conclude that the personalization based on grouping of patients achieved the highest sensitivity. In effect, the proposed methodology had better global behavior than models *M10* and *M12*, although it presented lower specificity than model *M22*. This aspect can be directly explained based on the criteria defined in

Figure 3.9. Actually, according to the clinical usefulness concept, the implemented criteria favored sensitivity over specificity.

Dataset	%	*M10*	*M12*	*M22*	Groups
Testing dataset $n = 30$	SE	86.3 *(85.6; 86.9)*	87.3 *(86.6;87.9)*	80.6 *(79.9; 81.3)*	94.9 *(94.6; 95.3)*
	SP	44.1 *(42.5; 45.7)*	60.8 *(58.8;62.9)*	85.9 *(83.9; 86.4)*	71.1 *(69.3; 72.8)*
	Gmean	61.6 *(60.4; 62.7)*	72.8 *(71.5;74.0)*	82.8 *(82.1; 83.5)*	82.1 *(81.0; 83.1)*

Table 4.89 - Global assessment of the personalization strategy.

Some statistical significance tests were carried out to support the above mentioned conclusions.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	T	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	14.962	0.000	-24	58	0.000	-8.7	0.352	-9.40	-7.99
M10 vs. Groups				-24	43	0.000	-8.7	0.352	-9.40	-7.99
SE	ea	9.906	0.003	22	58	0.000	-7.7	0.345	-8.35	-6.96
M12 vs. Groups				22	45	0.000	-7.7	0.345	-8.35	-6.96
SE	ea	7.966	0.007	-37	58	0.000	-14.3	0.381	-15.1	-13.5
M22 vs. Groups				-37	43	0.000	-14.3	0.381	-15.1	-13.5

Table 4.90 - Sensitivity - Grouping vs. M10; M12; M22.

Based on the results presented in Table 4.90 it is possible to conclude that the highest mean value of sensitivity was obtained through the grouping strategy.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	T	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SP	ea	0.213	0.646	-23	58	0.000	-26.9	1.15	-29.2	-24.6
M10 vs. Groups				-23	57	0.000	-26.9	1.15	-29.2	-24.6
SP	ea	0.088	0.767	-7.7	58	0.000	-10.2	1.32	-12.9	-7.6
M12 vs. Groups				-7.7	56	0.000	-10.2	1.32	-12.9	-7.6
SP	ea	4.584	0.036	13	58	0.000	14.1	1.04	11.9	16.2
M22 vs. Groups				13	52	0.000	14.1	1.04	11.9	16.2

Table 4.91 - Specificity - Grouping vs. M10; M12; M22.

Table 4.91 shows that model *M22* achieved a higher specificity value than the proposed approach. This aspect could be easily circumvented through the proper adjustment of the selection criteria and/or by an alternative partition (clusters) of the data space.

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	T	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
Gmean	ea	0.855	0.359	-26	58	0.000	-20.5	0.762	-22.0	-18.9
M10 vs. Groups				-26	57	0.000	-20.5	0.762	-22.0	-18.9
Gmean	ea	0.125	0.725	-11	58	0.000	-9.33	0.793	-10.9	-7.73
M12 vs. Groups				-11	58	0.000	-9.33	0.793	-10.9	-7.73
Gmean	ea	5.301	0.025	1.1	58	0.241	0.720	0.607	-0.496	1.936
M22 vs. Groups				1.1	49	0.242	0.720	0.607	-0.501	1.940

Table 4.92 – Geometric Mean- Grouping vs. M10; M12; M22.

Finally, the analysis of geometric mean (Table 4.92) showed that the proposed methodology achieved a higher value than *M10* and *M12*, and is equivalent to the one obtained in model *M22*.

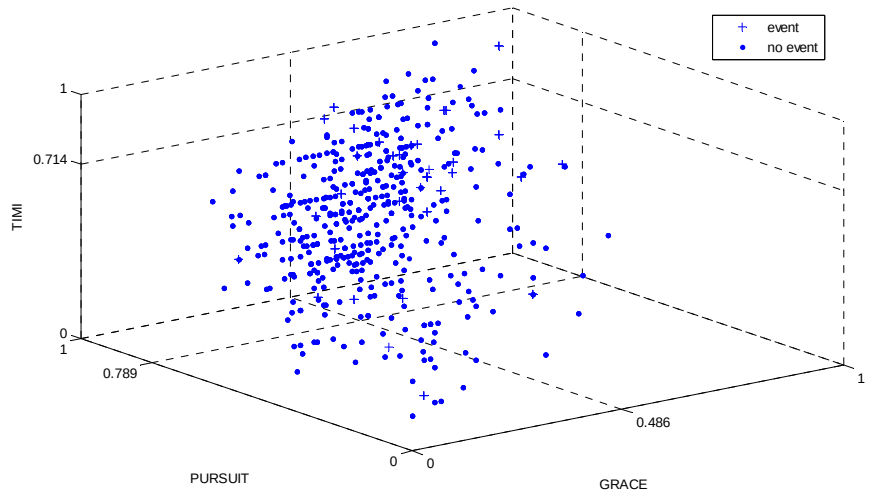
4.5.2 Tools Applied in Clinical Practice

This validation scenario considered the risk assessment tools that are applied in the daily clinical practice and it involved a real patient testing data set to provide the *true data*.

After the selection of the risk assessment tools, the first step of the proposed personalization strategy (Figure 3.2) was the dimensionality reduction. As referred, the high number of risk factors along with their heterogeneity imposed a dimensionality reduction step that in this case was applied to patients from Santa Cruz Hospital dataset (combined endpoint)[124].

[124] The other two datasets (Santa Cruz dataset, Santo André) are severely imbalanced with a reduced number of events. This limitation obstructed their incorporation in the validation procedure.

The dataset after the dimensionality reduction is presented in Figure 4.8. This dataset was obtained through the reduction of the original $p = 16$ risk factors to the $Q = 3$ outputs of the selected risk assessment tools[125].



$$\mathbf{y}_i = [\,y_R^i \;\; y_P^i \;\; y_T^i\,];\; c_R^i = \begin{cases} 0;\; y_R^i \le 0.486 \\ 1;\; y_R^i > 0.486 \end{cases};\; c_P^i = \begin{cases} 0;\; y_P^i \le 0.789 \\ 1;\; y_P^i > 0.789 \end{cases};\; c_T^i = \begin{cases} 0;\; y_T^i \le 0.714 \\ 1;\; y_T^i > 0.714 \end{cases}$$

Figure 4.8 - Dimensionality reduction.

Thus, the dimensionality reduction procedure mapped the original dataset $\mathbf{X}_{16\times460}$ on a low dimensional space $\mathbf{Y}_{3\times460}$ where each patient is characterized by the outputs of each one of the considered risk assessment tools.

The subtractive clustering algorithm was applied to $\mathbf{Y}_{3\times460}$, originating $K = 23$ clusters.

Table 4.93 presents the SE/SP values obtained for each risk assessment tool in each cluster.

[125] In this case the dimensionality reduction is based on the outputs of selected current risk assessment tools (g**R**ace, **P**ursuit and **T**imi).

C	GRACE		PURSUIT		TIMI		P	E
	SE	SP	SE	SP	SE	SP		
1	0	100	0	96.7	0	100	31	1
2	0	100	0	100	0	100	34	1
3	80	26.9	40	46.2	100	0	31	5
4	100	25	25	25	0	100	24	4
5	0	100	0	100	0	100	20	0
6	0	100	0	100	0	95	20	0
7	100	95.8	0	100	100	0	25	1
8	0	90.5	0	76.2	0	100	21	0
9	0	100	0	84.2	0	89.5	19	0
10	0	84.6	100	23.1	0	100	14	1
11	100	0	100	5.6	100	0	21	3
12	0	15	100	35	100	0	22	2
13	0	100	0	100	0	100	14	2
14	0	100	0	31.3	0	0	16	0
15	0	64.3	0	100	0	100	15	1
16	0	100	0	100	0	100	26	1
17	0	100	0	93.8	0	100	17	1
18	100	0	66.7	20	0	100	21	6
19	0	75	0	100	0	100	12	0
20	0	90.9	0	100	0	100	12	1
21	0	100	0	100	0	100	12	0
22	100	0	50	11.1	0	100	11	2
23	0	100	0	100	0	71.4	22	1

Table 4.93 - Performance of selected individual risk assessment tools in each cluster.

Bootstrapping validation ($N_B = 1000$) was applied to the original testing dataset. The low event rate obstructed the implementation of an alternative validation strategy, e.g. cross validation,

For each bootstrap sample $D_B = \{(\mathbf{x}_1, c_1)...,(\mathbf{x}_N, c_N)\}$, $N = 460$, each instance $\mathbf{x}_i \in \Upsilon_B$ was assigned to a specific cluster and it was classified by the individual model that best classifies the patients that belong to that cluster. The assessment of the grouping strategy as well as the individual risk assessment tools in each sample was performed considering the *true data* provided by that bootstrap sample. Table 4.94 presents the obtained results:

Dataset	%	GRACE	PURSUIT	TIMI	Groups
Bootstrap Samples $N_B = 1000$	SE	60.8 *(60.2; 61.3)*	42.4 *(41.9;43.1)*	33.5 *(33.0; 34.0)*	72.9 *(72.6; 73.5)*
	SP	74.9 *(74.8; 75.1)*	74.2 *(74.1;74.3)*	73.6 *(73.5; 73.7)*	75.1 *(75.0; 75.2)*
	Gmean	67.3 *(67.0; 67.6)*	55.8 *(55.5;56.2)*	49.3 *(48.9; 49.7)*	73.9 *(73.7; 74.4)*

Table 4.94 - Global assessment of the personalization strategy.

It is possible to conclude that the proposed combination of risk assessment tools reached a higher sensitivity than all the individual tools (the best individual sensitivity is 60.8% while the sensitivity for the proposed strategy is 72.9%). The specificity values are equivalent among the several models (the best individual specificity 74.9% equals the value obtained through the proposed strategy).

Statistical significance tests (Student's t-test) were applied to compare the performance of the personalization strategy (groups) and the best individual risk assessment tool (GRACE). The following table presents the obtained results:

		Levene's Test equal. Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% CI of the difference lower	Upper
SE	ea	5.09	0.024	32.1	1998	0.000	11.5	0.36	10.9	12.3
Groups *vs* GRACE				32.1	1983	0.000	11.5	0.36	10.9	12.3
SP	ea	2.45	0.115	-0.67	1998	0.498	-0.06	0.09	-0.25	0.12
Groups *vs* GRACE				-0.67	1992	0.498	-0.06	0.09	-0.25	0.12
Gmean	ea	17.8	0.000	30.9	1998	0.000	6.18	0.19	5.79	6.87
Groups *vs* GRACE				30.9	1955	0.000	6.18	0.19	5.79	6.87

Table 4.95 - Groups vs. GRACE [Santa Cruz dataset (endpoint: death/myocardial infarction)].

The outcomes of the test (Table 4.95) confirmed that the personalization strategy had a higher sensitivity than the best individual risk assessment tool. This improvement did not originate the reduction of the specificity value which is reflected in the geometric mean's value.

4.6 Conclusions

Based on global results derived in Section 4.2 (simulation), it seems plausible to affirm that the proposed combination methodology has potential to improve the risk prediction. The analysis of the ability to deal with missing risk factors also showed that global Bayesian model had a better risk assessment performance than the other two classifiers. This confirms its inherent ability to deal with missing risk factors.

The results[126] obtained with the combination of risk assessment tools applied in clinical practice (Section 4.3) are in accordance with these conclusions. The global Bayesian model presented better performance than the individual risk assessment tools in several test cases. Moreover, the optimization based on genetic algorithms improved the SE/SP values of risk prediction. However, in some test cases the optimization procedure did not improve the specificity values obtained with the individual risk assessment tools. The analysis of the classifiers' performance showed that the Bayesian global model after optimization had the best behavior when dealing with missing risk factors. In this validation procedure, bootstrapping validation was implemented due to some restrictions of the available testing datasets.

The incorporation of additional clinical knowledge[127] was assessed in Section 4.4. The obtained results reveal that the incorporation of additional clinical knowledge may have a positive impact in the risk prediction. The bootstrapping validation was also applied to reinforce the obtained results.

As explained, an additional methodology was developed to address the problem of the eventual lack of performance exhibited by CVD risk assessment tools. The results obtained through the implementation of the personalization strategy confirm that it is possible to achieve higher sensitivity without reducing the specificity values. Section 4.5 contains the obtained results with the Santa Cruz Hospital dataset (combined endpoint)[128]. Bootstrapping was also applied to support the derived results.

[126] Similarly to the previous situation, all the results were supported with statistical significance tests.

[127] The outcome of the incorporation of Body Mass Index (BMI) was evaluated.

[128] The low event rate (low number of events) in the other datasets inhibited their inclusion in the validation procedure.

5. Final Considerations

5.1 Introduction

The main motivation of this work is to provide a valid contribution for the improvement of the prediction (risk assessment) of a cardiovascular event, namely:

- To consider the available knowledge. Rather than to derive a new model, the proposed approach combines current CVD risk assessment tools;
- To avoid the need of choosing a risk assessment tool as a standard tool, this work allows the selection of one or more tools to make the risk assessment;
- To make possible the consideration of a higher number of risk factors;
- To cpe with missing information (missing risk factors);
- To enable tohe incorporation of empirical clinical knowledge (new risk factors) that physicians decide should be ideally integrated;
- To assure the clinical interpretability of the model, i.e. the capability of the model to express the behavior of the system in an understandable (clinical perspective) way;
- To improve the performance of the risk assessment when compared to the one achieved by the current risk assessment tools.

In order to accomplish these goals, two methodologies were developed: *i*) combination of individual risk assessment tools; *ii*) personalization based on grouping of patients.

The results obtained based on the combination of individual risk assessment tools are discussed in Section 5.2. Two different validation scenarios were applied to two specific patient conditions: *i*) heart failure patients; *ii*) coronary artery disease patients. The incorporation of clinical knowledge, in particular the results obtained with the BMI, is also discussed in this section.

Section 5.3 includes the discussion of the results obtained with the personalization based on grouping of patients' methodology. The validation of this approach was performed considering the same validation scenarios that were applied in the validation of the combination methodology.

This thesis presents some important contributions to the improvement of risk prediction. However, the author is aware that some additional research should be performed in order to complement and further improve the developed methodologies. The ongoing research issue is addressed in Section 5.4.

Finally, the scientific publications produced during the elaboration of this thesis are enumerated in Section 5.5.

5.2 Combination Methodology

5.2.1 Heart Failure

The first validation scenario was related with the combination of current risk assessment tools applied to one-year death risk assessment in heart failure patients. There are several risk assessment tools specific to heart failure patients (Table 2.2) that might be combined. However, the restrictions of the available dataset directly influenced the validation procedure, i.e. the selection of the current tools. Actually, the limitation of the TEN-HMS did not enable the direct use of current risk tools, (e.g. Seattle Heart Failure Model, the ABC Heart Failure Score, etc.). In order to circumvent this additional difficulty, simulated models were derived and afterwards combined. Thus, it is important to emphasize that in this case the developed combination scheme was exclusively validated with simulated models.

The flexible framework that was originated through the combination methodology was expected to overcome some of the identified drawbacks of existing risk assessment tools, such as: *i*) the discarding of the information provided by other tools; *ii*) limited number of risk factors used by each individual tool; *iii*) selection of a standard tool to apply in the clinical practice; *iv*) inability to deal with missing risk factors. Additionally, the risk prediction performance should be enhanced or at least maintained in relation to the one obtained with the individual models.

The obtained results suggest the validity of the proposed combination scheme. The global model created through the proposed combination methodology had higher accuracy and sensitivity than those of the individual models that were combined. The mean value of the individual models' specificity was higher than the one obtained with the global model. This aspect represents an undesired effect of the combination scheme that should be further investigated[129].

The ability of the developed strategy to cope with missing risk factors was also assessed. Bayesian inference mechanism had better performance and lower error's variance than the Cox regression model, meaning that the former is less vulnerable to missing input information.

These conclusions are supported by the results described in Table 4.7 and 4.17. It is important to emphasize that this validation process was significantly influenced by the limitations of the available dataset.

5.2.2 Coronary Artery Disease

The second validation scenario refers to the combination of current tools applied to a thirty day event[130] risk assessment in coronary artery disease patients[131]. In this case the available testing datasets allowed the combination of current risk assessment tools (Table 4.23) that are applied in the daily clinical practice.

The first procedure relied on the calibration/adjustment of the selected risk assessment tools considering the two specific populations/testing datasets[132]. Several experiments were performed which originated some important conclusions: *i*) the adjustment of individual tools' output categories must be carried out as presented in Figure 4.2; *ii*) the original calibration should be adopted in spite of the modest behavior of the risk assessment tools in the population under analysis; *iii*) GRACE risk assessment tool presented the best performance. TIMI and PURSUIT are not well adapted to the particular characteristics of the tested Portuguese populations. This evidence reveals that the particular characteristics of a population may have a

[129] The availability of additional testing datasets would be very important to enable a deeper insight on this issue.

[130] Death; Myocardial Infarction

[131] NSTEMI patients

[132] Testing datasets made available by Santa Cruz hospital (Lisbon) and Santo André hospital (Leiria).

great impact in the performance of a tool that was statistically derived (e.g. Cox regression) from a different population. Accordingly, different risk assessment tools may have dissimilar behaviors when applied to the same population as well as a specific tool may perform diversely among different populations.

Similarly to the previous validation scenario (simulation), the Bayesian global model was built based on the weighted average combination approach. An overall assessment was carried out through the comparison of the Bayesian global model's performance with the performances achieved by the individual risk assessment tools as well as by a voting approach. The combination methodology seems very interesting as it creates a Bayesian model that joins a reasonable performance with a set of important advantages when compared with classical statistical approaches: *i*) it creates a flexible framework since it allows the integration of knowledge from several sources (risk assessment tools); *ii*) it permits the integration of a higher number of risk factors in the risk assessment; *iii*) it avoids the need of selecting a specific model as a standard model in the clinical practice.

An additional optimization, based on genetic algorithms, was performed to adjust the global model to the specific populations under analysis. In the majority of the testing situations this procedure increased the specificity/sensitivity of the Bayesian global model which achieved a better performance than the remaining models/tools. This possibility of being easily adjusted to a specific population through the optimization of its parameters is an additional aspect that confirms the flexibility of the proposed approach.

The Bayesian global model after optimization also registered the best performance when there were missing risk factors, which suggests that the Bayesian inference mechanism is more suitable than the other tested models to deal with missing risk factors. The ability of coping with missing risk factors is another important advantage of the Bayesian approach. Furthermore, this feature also contributes to its adaptation to specific conditions, e.g. if a physician considers that a specific variable is not important for that population (no cardiac arrest admission situations), the model can be created without that specific variable.

Tables 4.30; 4.32; 4.36; 4.37; 4.38 and 4.42 are particularly important to support the above mentioned conclusions.

5.2.3 Incorporation of Clinical Knowledge

The possibility to incorporate additional clinical knowledge in risk assessment is one of the main goals of this work. The cardiologists that collaborated in this work identified the body mass index (BMI) as a risk factor that should be integrated in the risk assessment. Therefore, the influence of the integration of BMI in the risk prediction (combination scheme) was directly assessed.

As expected the incorporation of one additional risk factor did not originate a very significant improvement on the performance of the different models (Table 4.84). Although, the obtained results also demonstrated that the integration of a new risk factor can originate an improvement of the risk assessment.

This conclusion is important as the proposed combination approach easily allows the integration of new risk factors which can originate a positive effect in the model's performance (improve the risk prediction).

Ideally, this analysis should be extended to other risk factors and validated with other datasets. Nonetheless the unavailability of proper data did not allow that desired additional validation.

5.2.4 Conclusions

Considering the previous analysis that are supported by the results presented in Sections 4.2 and 4.3 it is possible to state that the proposed combination scheme reasonably achieved the initial targets of this work, as it allows:

- To consider the available knowledge provided by previously developed tools;
- To avoid the need of choosing a risk assessment tool as a standard tool;
- To consider a higher number of risk factors;
- To cope with missing risk factors;
- To enable the incorporation of new risk factors;
- To assure the clinical interpretability of the model.

This evidence makes this work a valid contribution for the improvement of the risk assessment applied to cardiovascular diseases.

The optimization procedure significantly improved the performance of the Bayesian global model. However, there are some test cases where the adopted metrics' values, namely the specificity value, were lower than values obtained by

some of the individual risk assessment tools. This aspect is directly related with the implemented optimization procedure, *multiobjective optimization*, where a tradeoff between objectives (maximize sensitivity/maximize specificity) must be found. Actually, the criteria for the selection of possible solutions favored sensitivity[133] which explains the eventual problems with the specificity values. In spite of an extensive set of experiments/iterations it was not possible to solve this apparent lack of performance related in particular with the specificity value. In this context, a personalization strategy was developed to minimize this weakness.

5.3 Personalization based on Grouping of Patients

The main objective of the personalization strategy based on grouping of patients was the creation of a methodology that can assure a better risk prediction than the one achieved by the current risk assessment tools.

The validation of this methodology also involved two scenarios that were applied to different patient conditions: *i*) heart failure; *ii*) coronary artery disease.

5.3.1 Heart Failure

Initially, this approach was validated exclusively based on the simulated models derived in Section 4.2. After several experiments, a set of simulated individual models were selected to implement the proposed personalization based on grouping of patients.

The obtained results with this first validation scenario (Table 4.89) appeared very encouraging as the proposed approach achieved the highest sensitivity also assuring a high geometric mean value. Although one of the selected simulated individual models presented a higher specificity value than the grouping strategy. This aspect was originated by the particular conditions of the simulation procedure and could be easily bypassed through the proper adjustment of the selection criteria and/or by an alternative partition (clusters) of the data space.

Therefore, it is possible to affirm that this first validation procedure suggests the potential of the proposed methodology to enhance the risk prediction performance.

[133] According to clinical usefulness explained in Section 4.3.

5.3.2 Coronary Artery Disease

This second validation scenario considered the grouping of patients based on three current risk assessment tools[134] applied in the daily clinical practice to coronary artery disease patients.

Based on the obtained results with this second validation scenario (Table 4.94) it was possible to confirm that the main goal of the proposed methodology was accomplished. The implemented personalization strategy allowed higher sensitivity values than the individual risk assessment tools without reducing the specificity values. Considering the high heterogeneity of original data this methodology obtained very reasonable results.

Nonetheless, the specific characteristics of the testing datasets (reduced number of patients, high heterogeneity of original data) highly influenced the validation procedure. A deeper validation would require a larger and more balanced dataset.

5.4 Ongoing Research

The proposed methodologies configure valid contributions to the improvement of CVD risk assessment. However, the author is aware that some research should be extended to improve the developed methodologies and respective results.

Regardless of the new developments additional testing datasets will be required to assure a consistent validation. In this context, the collaboration with clinical partners in order to obtain additional datasets must be the main focus of the ongoing research. Moreover, the possibility of implementing a prospective study to collect some specific data seems also very useful[135]. It is important to emphasize that the datasets that were used in this work were obtained as a result of a long process of collaboration with some hospitals[136].

[134] GRACE, PURSUIT, TIMI.

[135] This possibility is almost confirmed through the collaboration with Santo André hospital.

[136] Castle Hill hospital, Hull, UK; Santo André, Leiria, Portugal; Santa Cruz, Lisbon, Portugal.

As mentioned, the main limitation of this work is related with the unavailability/difficulty to obtain proper testing real patient datasets which somewhat negatively influenced the validation procedure.

The availability of data will trigger some of the following possibilities for the development of the methodology:

- To improve the weighted average combination scheme. The proposed approach is able to cope with different weights of the individual models as well as to remove/incorporate a given risk factor. However, it does not permit the definition of weights[137] for the individual risk factors. This issue was identified by the main clinical partner[138] of this work:

  "The physician may have some difficulty to define the weights of the different individual models to combine (weighted combination). It is easier for physicians to define the weights/importance of individual variables."

  Probably this new feature would allow a better combination of the individual risk assessment tools having a positive impact in the performance of global risk prediction;
- To explore alternative classifier structures (semi naïve Bayes classifiers) to implement the common representation of individual risk assessment tools. This alternative should be considered, even so it may impose an additional difficulty to the clinical interpretability of the model;
- To improve the optimization procedure using genetic algorithms or an alternative optimization technique. The main goal would be to obtain better performances considering a lower neighborhood[139] of the initial conditional probabilities table values;
- To improve personalization. In this thesis two separate methodologies were proposed to minimize the identified weaknesses of the current risk assessment tools. The creation of an approach that merges the two methodologies seems very interesting as it will gather the advantages of both approaches. Simultaneously, additional dimensionality reduction techniques (e.g. multi-output regularized feature projection *(Yu, 2006)*, etc.) should be tested in

[137] Different from values: 0 (removal); 1(incorporation).

[138] Dr. João Morais from Leiria-Pombal Hospital Centre.

[139] Assure the clinical significance (interpretability) of the model.

order to enhance the grouping of patients. The personalization can have a critical role in the motivation of the individual patient;

- To capture the dynamics of the risk evolution. The evaluation of the risk's evolution as a direct consequence of the risk factors' modification is an important feature to improve the risk prediction. Similarly to personalization, the correct assessment of risk dynamics can be essential to the motivation and self-responsibility of the individual patient.

5.5 Scientific Publications

Throughout this research some scientific publications were produced with the main aim of presenting the developed work along with the obtained results. These papers can be organized as: *i*) international conferences; *ii*) scientific journals.

5.5.1 International Conferences

- S. Paredes, T. Rocha, P. de Carvalho, J. Henriques, J. Morais, J. Ferreira, M. Mendes, "Cardiovascular Event Risk Assessment – Fusion of Individual Risk Assessment Tools Applied to the Portuguese Population", *15th International Conference on Information Fusion, Singapore, 2012;*
- S. Paredes, T. Rocha, P. de Carvalho, J. Henriques, J. Morais, J. Ferreira, M. Mendes, "Improvement of CVD Risk Assessment Tools' performance through innovative Patients' Grouping Strategies", *34th Annual International IEEE EMBS Conference, San Diego, 2012*;
- S. Paredes, T. Rocha, P. Carvalho, J. Henriques, "Aplicação de Algoritmos Genéticos na Optimização de Probabilidades Condicionais em Modelos Bayesianos de Risco Cardiovascular", *Congresso de Métodos Numéricos em Engenharia - CMNE 2011, Coimbra, 2011*;
- S. Paredes, T. Rocha, P. de Carvalho, J. Henriques, D. Rasteiro, J. Morais, J. Ferreira, M. Mendes,"Fusion of Risk Assessment Models with application to Coronary Artery Disease Patients ", *33th Annual International IEEE EMBS Conference, Boston, 2011*;

- S. Paredes, T. Rocha, P. Carvalho, J. Henriques, M. Harris, J. Morais "Cardiovascular Risk and Status Assessment*", 32th Annual International IEEE EMBS Conference, Argentina, 2010*;
- S. Paredes, T. Rocha, P. Carvalho, J. Henriques, M. Harris, J. Morais "Long Term Cardiovascular Risk Models' Combination - A new approach", *31th Annual International IEEE EMBS Conference, USA, 2009.*

5.5.2 Scientific Journals

- S. Paredes, T. Rocha, P. de Carvalho, J. Henriques, D. Rasteiro, J. Morais " Integration of Different Models to Improve the Death Risk Assessment in Heart Failure Patients – A Simulation Study " (*submitted to Computers in Biology and Medicine Journal in October 2011; awaiting submission results*);
- S. Paredes, T. Rocha, P. Carvalho, J. Henriques, M. Harris, J. Morais, "Long Term Cardiovascular Risk Models' Combination", *Computer Methods and Programs in Biomedicine Journal, March 2011.*

References

Aaronson, K. (1997). Development and Prospective Validation of a Clinical Index to Predict Survival in Ambulatory Patients Referred for Cardiac Transplant Evaluation. *Journal of American Heart Association*, Vol. 95 pp. 2660-2667.

Adlam, D. (2005). Using BNP to develop a risk score for heart failure in primary care. *European Heart Journal, Oxford Journals*, Vol. 26, pp. 1086–1093.

Aha, D. (1991). Instance-Based Learning Algorithms. *Machine Learning*, Vol. 6: pp. 37–66.

Allwein, E. (2000). Reducing multiclass to binary: a unifying approach for margin classifiers. *Proceedings of the 17th International Conference on Machine Learning*, pp. 9-16.

Alpert, J. (2000). Myocardial infarction redefined--a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. *Journal of the American College of Cardiology* , Vol. 36, pp. 959-969.

Alty, S. (2007). Predicting Arterial Stiffness from the Digital Volume Pulse Waveform. *IEEE Transactions on Biomedical Engineering*, Vol. 54, n. 12.

Andristos, P. (2002). *Data Clustering Tecnhiques, .* Technical Report CSRG-443, Department of Computer Science, University of Toronto.

Anile, A. (2005). Comparison among Evolutionary Algorithms and Classical Optimization Methods for Circuit Design Problems. *Proceedings of Evolutionary Computation Congress.*

Antman, E. (2000). The TIMI risk score for Unstable Angina / Non-St Elevation MI – A method for Prognostication and Therapeutic Decision Making. *Journal of American Medical Association- JAMA*, Vol. 284, n°7, pp. 835-842.

Assmann, G. (2002). Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular

Münster (PROCAM). *Circulation, AHA - American Heart Association*, Vol.105 pp. 310-315.

Ata, N. (2007). Cox Regression Models With Nonproportional Hazards Applied to Lung Cancer Survival Data, in,. *Hacettepe Journal of Mathematics and Statistics*, Vol. 36, pp. 157-167.

Atoui, H. (2006). Cardiovascular Risk stratification in decision support systems: A probabilistic approach. Application to pHealth. *Computers in Cardiology*, Vol. 33, pp. 281-284.

Ayers, S. (2007). *Cambridge Handbook of Psychology, Health and Medicine.* ISBN: 978-0521605106, Cambridge University Press.

Bäck, T. (1995). Generalized Convergence Models for Tournament Selection. *Proceedings of the 6th International Conference on Genetic Algorithms.* Morgan Kaufmann Publishers Inc.

Balasubramanian, V. (2009). Support Vector Machine Based Conformal Predictors for Risk of Complications following a Coronary Drug Eluting Stent Procedure. *Computers in Cardiology*, Vol. 36, pp. 5-8.

Bauer, E. (1998). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants. *Machine Learning*, Vol. 36, pp. 1-38.

Bertrand, M. (2002). Management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. *European Heart Journal, Oxford Journals*, Vol. 23, pp. 1809–1840.

Bidgoli, B. (2004). A comparison of resampling methods for clustering ensembles. *International Conference on Machine Learning, Models, Technologies and Applications*, pp. 939-945.

Boersma, E. (2000). Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation; Results from an international trial of 9461 patients. *Circulation, American Heart Association - AHA*, Vol. 101 pp. 2557-2657.

Boutilier, C. (1996). Context - Specific Independence in Bayesian Networks. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pp.115--123.

Bouvy, M. (2003). Predicting mortality in patients with heart failure: a pragmatic approach. *Heart, BMJ*, Vol. 89, pp. 605-609.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, Vol. 24, pp. 123-140.

Breiman, L. (2001). *Random Forests.* Statistics Department, University of California.

Brighton, H. (2002). Advances in Instance Selection for Instance-Based Learning Algorithms. *Data Mining and Knowledge Discovery*, Vol. 6: pp. 153–172.

Brophy, J. (2004). A Multivariate Model for Predicting Mortality in Patients with Heart Failure and Systolic Dysfunction. *The American Journal of Medicine*, Vol. 116.

Burges, C. (1998). A tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, Vol. 2: pp. 121–167.

Burton, A. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer*, Vol.91. pp 4-8.

Butler, J. (2008). Incident Heart Failure Prediction in the Elderly: The Health ABC Heart Failure Score. *Circulation, AHA - American Heart Association*, Vol. 105, pp. 310-315.

Campbell, C. (2002). Kernel methods: a survey of current techniques. *Neurocomputing*, Vol. 48, pp. 63-84.

Carmo, I. (2006). Prevalence of obesity in Portugal. *Obesity reviews*, Vol. 7, pp. 233-237.

CEC/EU (2005). *Confronting demographic change: a new solidarity between the generations - Green paper.* Commission of the European Communities. Accessed in December 2010: http://eur-lex.europa.eu/LexUriServ .

Choi, S. (2010). A Survey of Binary Similarity and Distance Measures. *Cybernetics and Informatics*, Vol.8 pp. 43-48.

Cleland, J. (2005). Noninvasive Home Telemonitoring for Patients With Heart Failure at High Risk of Recurrent Admission and Death. *Journal of American College of Cardiology*, Vol. 45, pp. 1654-1664.

Conroy, R. (2003). Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European Heart Journal, Oxford Journals*, Vol. 24, pp. 987-1003.

Cooper, F. (1999). An Overview of the Representation and Discovery of Causal Relationships using Bayesian Networks, in Computation, Causation and Discovery, AAAI Press and MIT Press. *Computation, Causation and Discovery, AAAI Press and MIT Press.*

Cordella, L. (1999). Reliability Parameters to Improve Combination Strategies in Multi-Expert Systems. *Pattern Analysis & Applications*, Vol. 2, pp. 205-214.

Cortes, C. (1995). Support-Vector Networks. *Machine Learning*, Vol. 20: pp. 273–297.

Cox, J. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *British Medical Journal*, Vol. 136.

D'Agostino, R. (2008). General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation, AHA - American Heart Association*, Vol. 117 pp. 743-757.

Davison, A. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

Davison, A. (2006). *Bootstrap Methods and their Apllication.* Cambridge University.

Demuth, H. (2002). *Neural Network Toolbox for Use with MATLAB. User's Guide.* MathWorks.

Domingos, P. (1996). Beyond Independence: Conditions for the optimality of the simple Bayesian classifier. *Proceedings of the 13th International Conference of Machine Learning*, pp. 105-112.

Dowdy, S. (2004). *Statistics for Research.* ISBN: 978-0471267355, Wiley-Interscience; 3rd edition.

Dreiseitl, S. (2005). Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artificial Intelligence in Medicine, Elsevier,* Vol. 33, pp. 25-30.

Duda, R. (2000). *Pattern Classification.* ISBN: 978-0471056690 Wiley-Interscience.

Dudina, A. (2011). Relationships between body mass index, cardiovascular mortality, and risk factors: a report from the SCORE investigators. *European Journal of Cardiovascular Prevention and Rehabilitation*, Vol. 18(5), pp.731-42.

Dwyer, K. (2007). *Decision Tree Instability and Active Learning, .* MSc Thesis submitted to the Faculty of Graduate Studies and Research, Department of Computing Science, University of Alberta.

EHN. (2008). *Healthy Hearts for All - Annual Report 2008.* European Heart Network . Accessed in December 2010: http://www.ehnheart.org/publications/annual-reports.html.

EHN. (2009). *Healthy Hearts for All - Annual Report 2009. .* European Heart Network.Accessed in December 2010: http://www.ehnheart.org/publications

Eiben, A. (2003). Introduction to Evolutionary Computing. ISBN: 978-3540401841, Springer.

Fletcher, R. (1999). *Practical Methods of Optimization 2nd Edition.* ISBN:0-471-91547-5, John Wiley & Sons.

Fodor, I. (2002). *A Survey on Reduction Techniques.* Lawrence Livermore National Laboratory .

Fonarow, G. (2005). Risk Stratification for In Hospital Mortality in Acutely Decompensated Heart Failure - Classification and Regression Tree Analysis. *Journal of American Medical Association*, Vol. 293 pp. 572-580.

Friedman, N. (1996). Lazy decision trees. *Proc. of the AAAI, MIT Press*, pp.717-724).

Friedman, N. (1997). Bayesian Network Classifiers. *Machine Learning*, Vol. 29, pp. 131—163.

Fung, G. (2001). *A Comprehensive Overview of Basic Clustering Techniques.* University of Wisconsin-Madison.

Furnkranz, J. (1999). Separate and Conquer Rule Learning. *Artificial Intelligence Review*, Vol. 13 pp. 3-54.

Gilli, M. (2008). A Review of Heuristic Optimization methods in econometrics. S*wiss Finance Institute Research* Paper Series, n. 12.

Gonzaléz, R. (2008). *Neural Networks for Variational Problems in Engineering.* Phd Thesis, Department of Computer Languages and Systems, University of Catalonia, Catalonia.

Graham, I. (2003). *Methods for Handling Missing Data. Research Methods in Psychology – Handbook of Psychology.* John Wiley & Sons. Inc.

Graham, I. (2007). Guidelines on preventing cardiovascular disease in clinical practice: executive summary. *European Heart Journal, Oxford Journals*, Vol.28, pp. 2375–2414.

Guyon, I. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* , Vol. 46 pp. 389-422.

Hammouda, K. (2000). *A Comparative Study of Data Clustering Techniques, SYDE 625: Tools of intelligent systems design.* Course Project, University of Waterloo.

Han, J. (2011). *Data Mining: Concepts and Techniques, 3rd edition.* ISBN: 978-0123814791 Morgan Kaufmann.

Heckerman, D. (1999). A tutorial on Learning with Bayesian Networks. *Learning in Graphical Models, MIT press.*

Hilton, A. (2006). Stat note 6. *Microbiologist*, pp. 34-36.

Hobbs, F. (2004). Cardiovascular disease: different strategies for primary and secondary prevention. *Heart, BMJ*, Vol. 90, pp. 1217–1223.

Hoeting, J. (1999). Bayesian Model Averaging: A Tutorial . *Statistical Science*, Vol. 14, pp. 382-417.

Hornik, K. (1991). Multilayer feedforward networks are universal approximators. *Neural Networks*, Vol. 4 pp. 251-257.

Horton, N. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Annual Statistics*, Vol. 61, pp. 79-90.

Hsu, C. (2003). Implications of the Dirichlet Assumption for Discretization of Continuous Variables in Naive Bayesian Classifiers. *Machine Learning*, Vol. 53, pp. 253-263.

IBM. (2010). *IBM SPSS Advanced Statistics 19.* IBM Company.

Jain, K. (1999). Data clustering: a review. *ACM Computing Surveys*, Vol. 31(3) pp. 264–323.

Janssen, K. (2009). Dealing with Missing Predictor Values When Applying Clinical Prediction Models. *Clinical Chemistry Journal*, Vol. 55, pp. 994-1001.

JBS. (2005). JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice. *Heart, BMJ*, Vol. 91, Supplement V.

Johnson, R. (2001). An Introduction to Bootstrap. *Teachning Statistics*, Vol. 23, pp. 49-54.

Jordan, M. (1998). *Learning in Graphical Models.* MIT Press.

Kannel, W. (1999). Profile for Estimating Risk of Heart Failure. *Archives of Internal Medicine, American Medical Association*, vol. 159. .

Kasamatsu, T. (2008). Application of support vector machine classifiers to preoperative risk stratification with myocardial perfusion scintigraphy. *Circulation Journal*, Vol. 72, pp. 1829-1835.

Keogh, E. (1999). Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches. *Proceedings of International Workshop on Artificial Intelligence and Statistics*, pp. 225-230.

Khanna, K. (2005). Missing medical information adversely affects care of patients. *British Medical Journal, BMJ*, Vol. 330.

Kirkwood, B. (2003). *Medical Statistics, 2nd Edition.* ISBN: 978-0-86542-871-2, Blackwell Science.

Kohavi, R. (1996). Scaling up the accuracy of naïve Bayes classifiers: a decision tree hybrid. *Proceedings of the 2nd International Conference Knowledge Discovery and Data Mining*, pp. 202-207.

Konak, A. (2006). Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering and System Safety* , Vol. 91, pp. 992–1007.

Koopman, R. (2008). Evaluating Multivariate Risk Scores for Clinical Decision Making . *Family Medicine*, Vol. 40 pp. 412-416.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, Vol.31, pp. 249-268.

Kubat, M. (1998). Machine Learning for the detection of oil spills in Satellite Radar Images. *Machine Learning*, Vol. 30, pp. 195-215.

Kutner, M. (2004). *Applied Linear Statistical Models, 5th edition.* ISBN: 978-0073108742, McGraw-Hill.

Langley, P. (1994). Induction of selective Bayesian classifiers, . *Proceedings of the 10th International Conference of Machine Learning*, pp. 399-406.

Lee, D. (2003). Predicting Mortality Among Patients Hospitalized for Heart Failure – Derivation and Validation of a Clinical Model. *Journal of American Medical Association- JAMA*, Vol. 290 n°19 pp. 258.

Lee, D. (2007). *Nonlinear Dimensionality Reduction.* ISBN: 978-0387393506 Springer.

Lee, K. (2008). *Modern Heuristic Optimization Techniques: Theory and Applications to Power Systems.* ISBN: 978-0471457114, Wiley - IEEE Press.

Levy, W. (2006). The Seattle Heart Failure Model Prediction of Survival in Heart Failure. *Journal of American Medical Association- AMA*, Vol. 113, pp. 1424-1433.

Logutov, O. (2005). Multi-model fusion and error parameter estimation. *Journal of the Royal Meteorological Society*, Vol. 131, pp. 3397–3408.

Lukas, L. (2003). *Least Squares Support Vector Machines Classification Applied to Brain Tumor Recognition Using Magnetic Resonance Spectroscopy.* ISBN 90-5682-460-0, K. U. Leuven.

Luong, T. (2003). A Comparison of the Performance of Classical Methods and Genetic Algorithms for Optimization Problems Involving Numerical Models. *Proceedings of Evolutionary Computation - CEC'03*, Vol. 3, pp. 2019-2025.

Maaten, L. (2009). Dimensionality Reduction: A Comparative Review. *TiCC TR 2009–005 Tilburg University.*

Mclachlan, G. (2008). *The EM Algorithm and Extensions; 2nd Edition.* ISBN: 978-0471201700, Wiley-Interscience.

Michalewicz, Z. (2004). *How to Solve it: Modern Heuristics 2nd Edition.* ISBN: 3-540-22494-7, Springer.

Mingers, J. (1989). An empirical comparison of selection methods for decision tree induction. *Machine Learning*, Vol.4 pp. 319-342.

Mitchell, T. (1997). *Machine Learning.* ISBN: 0-07-115467-1, McGraw-Hill.

Mocian, H. (2009). *Survey of Distributed Clustering Techniques.* MSc Idividual, Imperial College London.

Moons, K. (2009). Prognosis and prognostic research: what, why and how?. *British Medical Journal*, Vol. 338, pp. 1317-1320.

Morrow, D. (2008). TIMI Risk Score for ST-Elevation Myocardial Infarction: A Convenient, Bedside, Clinical Score for Risk Assessment at Presentation. *Circulation, American Heart Association*, Vol. 102; pp. 2031-2037.

Murthy, S. (1988). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery, SpringerLink*, Vol.2, pp. 345-389.

Neapolitan, R. (2004). *Learning Bayesian Networks.* ISBN: 0-13-012534-2, Pearson Prentice Hall.

Neocleous, C. (2002). Artificial Neural Network Learning: A Comparative Review. *Methods and Applications of Artificial Intelligence*, Vol. 2308, pp. 300-313.

Nicholson, A. (2008). *Decision Support for Clinical Cardiovascular Risk Assessment, Bayesian Networks - A Practical Guide to Applications.* ISBN: 978-0-470-06030-8, John Willey & Sons, England.

Niculescu, R. (2005). *Exploiting Parameter Domain Knowledge for Learning in Bayesian Networks.* School of Computer Science, Carnegie Mellon University.

Ning, G. (2006). Artificial neural network based model for cardiovascular risk stratification in hypertension. *Medical and Biomedical and Biological Engineering and computing*, Vol. 44, nº 3, pp. 202-208.

NVDPA. (2009). *National Vascular Disease Prevention Alliance. Guidelines for the assessment of absolute cardiovascular disease risk.* ISBN: 978-1-921226-38-0, National Heart Foundation of Australia.

Ordonez, C. (2006). Comparing Rules and Decision Trees for Disease Prediction. *Proceedings of International Conference of Information Knowledge and Management,* pp. 17-24.

Pan, W. (2006). Using Input Dependent Weights for Model Combination and Model Selection with multiple Sources of Data. *Statistica Sinica*, Vol. 16 pp. 523-540.

Parekh, R. (2000). Constructive Neural Network Learning Algorithms for Pattern Classification. *IEEE Transactions on Neural Networks*, Vol. 11(2), pp. 436-451.

Pazzani, M. (1996). Constructive Induction of Cartesian Product Attributes. *Information, Statistics and Induction in Science* , pp. 66-77.

Peddersen, M. (2010). *Tuning & Simplifying Heuristical Optimization, .* Phd Thesis, School of Engineering Sciences, University of Southampton.

Pitt, A. (2009). *Application of data mining techniques in the prediction of coronary artery disease: use of anaesthesia time-series and patient risk factor data.* Phd Thesis, Queensland University of Technology.

Pocock, S. (2005). Predictors of mortality and morbidity in patients with chronic heart failure. *European Heart Journal, Oxford Journals*, Vol. 27 pp-65-75.

Quinlan, J. (1992). *C4.5: Programs for Machine Learning.* ISBN: 978-1558602380, Morgan Kaufmann.

Raftery, A. (2003). *Using Bayesian Model Averaging to Calibrate Forecast Ensembles.* Department of Statistics, University of Washington.

Reiter, N. (2009). HeartCycle: Compliance and Effectiveness in HF and CAD Closed-Loop Management. *Proceedings of the 31st Annual International Conference of the IEEE EMBS.*

Ribeiro, B. (2008). Manifold Learning for Premature Ventricular Contraction Detection. *Computers in Cardiology,* pp. 917 - 920, IEEE.

Rich, M. (2006). Predicting survival in elderly patients with heart failure. *Archives of Internal Medicine - American Medical Association*, Vol. 166, pp. 1892-1898.

Roberts, J. (2006). *Bayesian Networks for Cardiovascular Monitoring.* Master thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Rossi, R. (2010). *Applied Biostatistics for the Health Sciences.* ISBN: 978-0-470-14764, John Wiley & Sons.

Samsa, G. (2005). Combining Information From Multiple Data Sources to Create Multivariable Risk Models: Illustration and Preliminary Assessment of a New Method. *Journal of Biomedical Biotechnology*, Vol.2 pp.113–123.

Schwenker, F. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *Proceedings of the 17th International Conference on Machine Learning*, pp. 9-16.

Senni, M. (2006). A novel Prognostic Index to Determine the Impact of Cardiac Conditions and Co-Morbidities on One-Year Outcome in Patients with Heart Failure, in. *The American Journal of Cardiology*, Vol. 98, pp. 1076-1082.

Shen, J. (2004). Adaptive Model Selection and Assessment for Exponential Family Distributions. *Technometrics*, Vol. 46 pp. 306-317.

Shen, J. (2007). *Using Cluster Analysis, Cluster Validation and Consensus Clustering to Identify Subtypes of Pervasive Developmental Disorders.* Master of Science Thesis, Queen's University, Canada.

Sheskin, D. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures: 3rd Edition .* ISBN: 978-1584884408, Chapman and Hall.

Sıma, J. (2003). General-purpose computation with neural networks: A survey of complexity theoretic results. *Neural Computation*, Vol.15 pp. 2727–2778.

Stevens, R. (2001). The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes. *Clinical Science, Portland Press*, Vol.101 pp. 671–679.

Steyerberg, W. (2009). *Clinical Prediction Models – A Practical Approach to Development. Validation and Updating.* ISBN: 978-0-387-77243-1, Statistics for Biology and Health, Springer.

Strehl, A. (2002). Cluster Ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, pp. 583 - 587.

Sugiyama, M. (2010). Semi-supervised local Fisher discriminant analysis for dimensionality. *Machine Learning*, Vol. 78 pp. 35-61.

Swedberg, K. (2005). Guidelines for the diagnosis and treatment of chronic heart failure: executive summary. *European Heart Journal, Oxford Journals*, Vol. 26, pp. 1115–1140.

Syed, A. (2011). *A Review of Cross Validation and Adaptive Model Selection.* Mathematics Theses, paper 99, Georgia State University.

Tang, E. (2007). Global Registry of Acute Coronary Events (GRACE) hospital discharge risk scores accurately predicts long term mortality post-acute coronary syndrome. *American Heart Journal*, Vol. 153, n° 1, pp. 30-35.

Todorovsky, L. (2003). Combining Classifiers with Meta Decision Trees. *Machine Learning*, Vol. 50, pp. 223–24.

Tsymbal, A. (2001). Ensemble feature selection with Dynamic Integration of Classifiers. *Proceedings of Congress on Computational Intelligence Methods and Applications, CIMA' 2001*, pp. 558-564.

Tsymbal, A. (2003). Ensemble feature selection with the simple Bayesian classification. *Information fusion*, Vol. 4, Issue.2, pp. 87-100.

Ture, M. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, Vol. 29, pp. 583-588.

Twardy, C. (2004). *Data mining cardiovascular bayesian networks.* Technical Report 2004/165, School of CSSE, Monash University.

Twardy, C. (2005). *Knowledge engineering cardiovascular Bayesian networks from the literature.* Technical Report 2005/170, Monash University.

Ulrich, J. (2008). *Supervised Machine Learning for Email Thread Summarization,.* MSc Thesis submitted to the Faculty of Graduate Studies, University of Brittish Columbia, Vancouver.

Valavanis, I. (2010). A multifactorial analysis of obesity as CVD risk factor: Use of neural network based methods in a nutrigenetics context. *BMC Bioinformatics*, Vol.11, pp. 453-463.

Vasquez, R. (2009). The MUSIC Risk score: a simple method for predicting mortality in ambulatory patients with chronic heart failure. *European Heart Journal, Oxford Journals*, Vol. 30, pp. 1088–1096.

Verduijn, M. (2007). Prognostic Bayesian networks II: An application in the domain of cardiac surgery. *Journal of Biomedical Informatics*, Vol. 40, pp. 619-630.

Visweswaran, S. (2007). *Learning Patient-Specific Models from Clinical Data. .* Phd thesis submitted to School of Arts and Sciences, University of Pittsburgh. .

Vivarelli, F. (2001). Comparing Bayesian neural network algorithms for classifying segmented outdoor images. *Neural Networks*, Vol. 14 pp. 427-437. .

Voss, R. (2002). Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks. *International Journal of Epidemiology*, Vol. 31, pp. 1253-1262.

Wallis, E. (2000). Coronary and cardiovascular risk estimation for primary prevention: validation of a new Sheffield table in the 1995 Scottish health survey population, in British Medica. *British Medical Journal*, Vol. 320 pp. 671-676.

Wang, H. (2006). Nearest Neighbors by Neighborhood Counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, nº. 6, pp. 942-953.

Wayman, J. (2003). *Multiple Imputation for Missing Data: What is and How Can I use It? .* Annual Meeting of the American Educational Research.

Webb, G. (2005). Not so naïve Bayes: Aggregating one-dependence estimators. *Machine Learning*, Vol. 58, pp. 5-24. .

Wehrens, R. (2000). The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, Vol. 54, pp. 35-52.

Wettschereck, D. (1997). A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. *Artificial Intelligence Review*, Vol. 11, pp. 273–314.

Whitley, D. (1995). Genetic Algorithms and Neural Networks. *Genetic Algorithms in Engineering and Computer Science*, pp. 191-201.

WHO . (2011). *BMI classification.* World Health Organization. Accessed on May 2011: http://apps.who.int/bmi/index.jsp?introPage=intro_3.html.

WHO. (2007). *Prevention of cardiovascular disease guidelines for assessment and management of cardiovascular risk.* World Health Organization. Accessed in December 2010: http://www.who.int/cardiovascular_diseases/resources.

WHO. (2009). *Cardiovascular Diseases (CVDs).* World Health Organization , Fact sheet n°317. Accessed in December 2010: http://www.who.int/mediacentre/.

Wilson, D. (2000). Reduction Techniques for Instance-Based Learning Algorithms. *Machine Learning*, Vol. 38, pp. 257-286. .

Witten, I. (2011). *Data Mining – Practical Machine Learning Tools and techniques, 3rd Edition.* ISBN: 978-0-12-374856, Morgan Kauffman.

Woodward, M. (2007). Adding social deprivation and family history to cardiovascular risk assessment – The ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart, BMJ*, Vol. 93, pp. 172-176.

Wormser, D. (2011). Body-mass index, abdominal adiposity, and cardiovascular risk. *The Lancet* , Vol. 378, Issue 9787, Page 228.

Xu, R. (2009). *Clustering.* ISBN: 978-0-470-27680-8, IEEE press, John Wiley & Sons Publications.

Yam, J. (2002). Feedforward Networks Training Speed Enhancement by Optimal Initialization of the Synaptic Coefficients. *IEEE Transactions on Neural Networks*, Vol.12; pp.430-434.

Yang, Y. (2002). A Comparative Study of Discretization Methods for Naïve-Bayes Classifiers. *Proceedings of Pacific Rim Knowledge Acquisition Workshop*, pp. 159-173.

Yang, Y. (2009). Discretization for naïve- Bayes learning managing discretization bias and variance. *Machine Learning*, Vol. 74, pp. 39-74.

Yen, S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, pp. 5718-5727.

Yu, S. (2006). Multi-Output Regularized Feature Projection. *IEEE Transactions on Knowledge and Data Engineering* , Vol. 18, nº 12.

Zhang, G. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, Part C 30 (4): pp. 451- 462.

Zhang, H. (2009). *Adaptive Model Selection in Linear Mixed Models.* University of Minnesota.

Zheng, Z. (2000). Lazy learning of Bayesian rules. *Machine Learning*, pp. 53-84.

Zheng, F. (2005). A comparative study of semi-naïve Bayes methods in classification learning. *Proceedings of the 4th Australasian Data Mining Conference*, pp. 141-156.

Zheng, F. (2006). Effcient Lazy Elimination for Averaged One-Dependence Estimators. *23rd International Conference on Machine Learning* , pp. 1113-1120.

Zheng, W. (2011). Association between Body-Mass Index and Risk of Death in More Than 1 Million Asians. *The New England Journal of Medicine*, Vol. 364, nº 8.