FRAMEWORK TO PREDICT THE METABOLIC SYNDROME WITHOUT DOING A BLOOD TEST

based on machine learning for a clinical decision support system

Mauricio Andres Barrios Barrios

Supervised by Miguel Jimeno, PhD and Pedro Villalba, PhD

Systems Engineering and Computing

Division of Engineering

Universidad del Norte

Nov, 2020

A dissertation submitted in partial fulfilment of the requirements for the degree of Ph.D. in Systems Engineering and Computing.

Declaration by Postgraduate Students

Authenticity of Dissertation

I hereby declare that I am the legitimate author of this Dissertation and that it is my original work.
No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education.
I hold the Universidad del Norte harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

Faculty/Institute/Centre/School	Division of Engineering
Degree	Ph.D. in Systems Engineering and Computing
Title	Framework to predict the metabolic syndrome without doing a blood test based on machine learning for a clinical decision support system
Candidate	Mauricio Andres Barrios Barrios
Signature of Student	_____
Date	February 19, 2021

To my beloved wife Belsy, and my lovely daughters Taliana and Gabriela

For their constant support and patience during the journey on this long and difficult road.

Acknowledgements

I immensely thank Jehovah God, his son Jesus Christ and the wonderful Holy Spirit for giving me the wisdom and the right path to carry out this life project.

To my wife Belsy Cotes, to my daughters Taliana Barrios, and Gabriela Barrios for their support and understanding every time I have to be away.

To my parents Jose Barrios Romero and Carmen Barrios Borrero for the education they gave me and instilled in me the love for Jesus Christ, my idol and respect for others.

I want to thank my brother Jose Barrios for taking care of me and advising me.

I also want to thank my in-laws Hernan Cotes and Linda Abdala for supporting me in this life project.

I want to thank my directors Miguel Jimeno and Pedro Villalba for trusting that I could lead this work and for allowing me to make my own decisions. I also want to thank Edgard Navarro for bringing all his expertise in metabolic syndrome as the base of this work, and supplying the dataset that are the object of this study

Finally, I want to express my deep thanks to the Administrative Department of Science, Technology, and Innovation—COLCIENCIAS of Colombia for the Doctoral scholarship. In addition, I also want to express my deep thanks to the Universidad del Norte and Universidad Autonoma del Caribe.

Abstract

Metabolic Syndrome (MetS) is a cluster of risk factors that increase the likelihood of heart disease and diabetes mellitus, and researchers have recently linked it to worse outcomes for the novel Covid-19 disease. It is crucial to get diagnosed with time to take preventive measures, especially for patients in locations without proper laboratories and medical consultations. This work presents a new model to diagnose metabolic syndrome using machine learning and non-biochemical variables that healthcare professionals can obtain from initial consultations. For evaluating and comparing the model, this work also proposes a new methodology for performing research on data mining called RAMAD. The methodology standardizes the novel model's comparison with similar classification models, using their reported variables and previously obtained data from a study in Colombia, using the holdout and random subsampling validation methods to get performance evaluation indicators between the models. The resulting ANN model used three hidden layers and only Hip Circumference, dichotomous Waist Circumference, and dichotomous blood pressure variables. It gave an Area under Receiver Operating Characteristic curves (AROC) of 87.75% by the International Diabetes Federation (IDF) and 85.12% by Harmonized Diagnosis or Joint Interim Statement (HMS) diagnosis criteria, higher than previous models. Thanks to the new methodology, diagnosis models can be thoroughly documented for appropriate future comparisons, thus benefiting the studied diseases' diagnosis.

Medical personnel needs to know the factors involved in the syndrome to start a treatment. So, this work also presents the segmentation of metabolic syndrome in types related to each biochemical variable. It uses the RAMAD methodology together with several machine learning techniques to design a framework to predict MetS and their several types, without using a blood test and only anthropometric and clinical information. The results showed an excellent system for predicting six MetS types that combine several factors mentioned above that have an AROC with a range of 71% to 96%, and an AROC 82.86%. This thesis finishes with the proposal of using a SCRUM Thinking framework for creating mobile health applications to implement the new models and serve as decision tools for healthcare professionals. The standard and fundamental characteristics were analyzed, finding the quality attributes verified in the framework's early stages.

Keywords — Metabolic Syndrome, Segmentation, Quine–McCluskey, Random Subsampling validation, RAMAD, Machine learning, Framework, International Diabetes Federation (IDF), Harmonized Diagnosis or Joint Interim Statement (HMS).

Contents

List of Figures

List of Tables

List of Abbreviations

GCP	Good Clinical Practices
WHO	World Health Organization
NCEP ATP III	National Cholesterol Education Programme Adult Treatment Panel III
EGIR	European Group for the study of Insulin Resistance
IDF	International Diabetes Federation
HMS	Harmonized Metabolic Syndrome
MetS	Metabolic Syndrome
CHD	Coronary Heart Disease
IR	Insulin Resistance
ICD	International Classification of Diseases
OR	Odds Ratio
CI	Confidence Interval
SS	Sensitivity
SP	Specificity
FNR	False Negative Rate
FPR	False Positive Rate
AROC	Area under Receiver Operating Characteristic Curve
WC	Waist Circumference
BP	Blood Pressure
HDL-C	High Density Lipoprotein Cholesterol
FPG	Fasting Plasma Glucose
TG	Triglycerides
WG	Weight
HG	Height
HC	Hip Circumference
WHR	Waist to Hip ratio
WSR	Waist to Stature
BMI	Body Mass Index
BFP	Body Fat Percentage
SBP	Systole Blood Pressure
DBP	Diastole Blood Pressure
ANN	Artificial Neural Networks
SMOTE	Synthetic Minority Over-sampling Technique
PCLR	Principal Component Logistic Regression
RUSBoost	Random Under Sampling Synthetic Minority Over-sampling Technique

1

Introduction

1.1 | Motivation from a biomedical perspective

Many people attend a medical check-up to determine their health status, but after the medical consultation, they only find weight gain due to eating habits and sedentary lifestyle. However, after proceeding to perform laboratory tests, the doctor discovers a set of metabolic disorders that include dyslipidemias (abnormal concentrations of lipids in the blood: increased triglycerides and decreased HDL cholesterol), hypertension, hyperglycemia, and obesity (Bruce and Byrne (2009)). These people suffer from Metabolic Syndrome (MetS), a series of metabolic risk factors that increase the probability of heart disease, a hemorrhagic cerebrovascular accident, or diabetes mellitus, where cardiovascular disease is one of the leading causes of morbidity and mortality worldwide (Kaur (2014)) and is the leading cause of death in Colombia (ONS (2013)).

This prevalence is why the academic community has cited this syndrome in more than 24,000 publications registered in PUBMED until May 2020. Of course, the number of publications or citations related to metabolic syndrome provides an estimate of the subject's importance as research in the scientific field. Worldwide, researchers recognize the metabolic syndrome as the trigger to increase the chances of developing heart disease and diabetes mellitus(Cornier et al. (2008); Tagle-Luzárraga et al. (2007)).

Even together with SAR-COVID19, it is a risk factor for accelerating the onset of a heart attack due to its relationship with obesity, hypertension, and blood glucose(Marhl et al. (2020)). Therefore, it is imperative to develop tools to diagnose metabolic syndrome early to avoid or reduce its consequences. Thanks to computer science, it is possible to determine the syndrome early.

1.2 | Motivation from a Computer Science perspective

One of the areas or fields of computer science with wide growth is machine learning, which studies the way machines learn through algorithms and allows the extraction of hidden patterns in a dataset(Marsland (2014)). Being one of the most common applications in machine learning is the use of decision making in different scenarios such as medical decision making(Cleophas and Zwinderman (2015)).

Medical decision making is an intellectual process that leads to a choice between several possibilities of a set of elements of judgment and actions carried out by health professionals based on their knowledge and experience (theirs or that of others) and the indications observed in the patient even further. Also, it is a continuous process that invariably affects the failure or success of the treatment. Since to make a diagnosis, the doctor has his clinical skills and knowledge as well as an orderly and efficient methodology to decide based on his diagnostic hypotheses, the treatment that leads to the solution of the patient's health problem(Alberto et al. (2010); Moncada (2013)). However, because it is a human decision, it presents errors and also depends on external processes (laboratory tests, diagnostic images) that generate latency when diagnosing a disease.

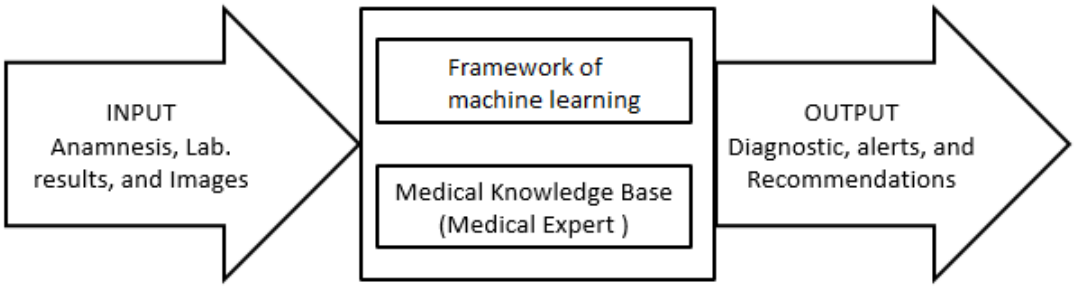

Figure 1.1: General model of a clinical decision support system.

It is important to note that clinical decision support system (CDSS), as shown in Figure 1.1 are computer systems designed to impact clinician decision making about individual patients at the point in time that these decisions are made(Berner and Gong (2016); Greenes (2014)) and the advantages and uses are multiple. One of the uses is to make a better decision when diagnosing health conditions. These systems base their core on an expert system's decision: a system that emulates the decision-making of one or more experts in the countryside. Another use is the decrease in the number of erroneous diagnoses, increasing the number of correct diagnoses, and improving the patient's quality of life. Likewise, it reduces public health spending and the number of visits to the doctor. A CDSS is possible thanks to machine learning frameworks inte-

grated as components because they rely on such frameworks to make decisions using logistic regression, decision trees, artificial neural networks, and others.

1.3 | Problem Statement

For this thesis, the focus will be on the delay or latency of the metabolic syndrome diagnosis because healthcare professionals require the lipid values and glucemy from the patient's blood. For doctors, having a Clinical Decision Support System (CDSS) is having one more support tool for their daily work. It is even the first element to consider in some cases. The doctor is the one who decides using his criteria, whether to follow the advice of the CDSS or not.

There are medical organizations made up of experts who have defined criteria to diagnose metabolic syndrome using five risk factors such as waist circumference, blood pressure, measurement of HDL-C biochemical variables, triglycerides, and blood glucose level (Navarro and Vargas (2012)). These criteria represent the basis of medical knowledge of a CDSS to diagnose the syndrome. However, a critical stage is missing is a framework of machine learning tools to diagnose metabolic syndrome without doing a blood test to avoid using biochemical variables and reduce the delay in its diagnosis.

1.4 | Objectives

1.4.1 | Main Objective

To design and implement a framework to predict the metabolic syndrome without doing a blood test based on machine learning

1.4.2 | Specific Objective

1. To analyze the prediction models of the metabolic syndrome published in the scientific community.
2. To develop and validate a model to predict the metabolic syndrome without doing a blood test sample.
3. To compare the proposed model with others found in the literature review.
4. To design and implement a framework to predict the metabolic syndrome without doing a blood test based on machine learning.

1.5 | Goals of the thesis

This thesis explains the solution to predicting metabolic syndrome early to avoid or delay the onset of some illnesses such as heart attack or/and diabetes mellitus. The thesis uses machine learning techniques and variables acquired using non-invasive methods, which require a blood test affecting the response time to treat the patient and start a treatment that reduces the risk factors for the consequent diseases.

For this purpose, this work proposes a methodology to design a new prediction model, beginning with a review of machine learning's technical principles for the prediction of metabolic syndrome published by the scientific community to analyze them. Then, I designed and developed with the helping of my supervisors a proposed new model to compare it based on the performance indicators such as AROC, sensitivity, and specificity.

Also, a framework for the implementation of the model was developed. This framework establishes new requirements for medical personnel that diagnose metabolic syndrome. Physicians always check the triglycerides, fasting blood glucose, and HDL-C values to recommend a specific treatment to prevent diseases such as diabetes or coronary diseases and know the possible cause since it allows a better decision to plan patient treatment (Alberti et al. (2009)).

Therefore, this thesis also details one solution for predicting metabolic syndrome using a non-invasive method without using biochemical variables and, at the same time, allows for identifying the possible cause that produces it.

1.6 | Contributions

The contributions of this thesis are:

- A methodology for the prediction of metabolic syndrome and similar health conditions
- A novel model to predict the metabolic syndrome using non-invasive methods without using biochemical variables
- A mathematical representation to diagnose metabolic syndrome using HMS criteria.
- The definition of metabolic syndrome types to help doctors design more accurate treatments.

- A framework for predicting the different types of metabolic syndrome, which obtained better results compared with the prediction of the traditional metabolic syndrome.

Such contributions are supported on the results here presented, and on the following publications:

1. Barrios, M., Jimeno, M., Villalba, P. (2018). Minería de datos aplicada a la investigación del síndrome metabólico. In Desarrollo tecnológico e Innovación. Ed: Coruniamericana ISBN: 978-958-5512-34-4

2. Barrios, M., Jimeno, M., Villalba, P., Navarro, E. (2019). Novel Data Mining Methodology for Healthcare Applied to a New Model to Diagnose Metabolic Syndrome without a Blood Test. Diagnostics, 9(4), 192. https://doi.org/10.3390/diagnostics9040192. (SJR Q2).

3. Barrios, M., Jimeno, M., Villalba, P. Scrum Thinking: A Framework for the development of mHealth. In Proceedings of the 11th International Multi-Conferences on Complexity, Informatics and Cybernetics: IMCIC 2020, Orlando, Florida, USA, 10-13 March 2020. (Scopus)

4. Barrios, M., Jimeno, M., Villalba, P., Navarro, E. (2020). Framework to Diagnose the Metabolic Syndrome Types without using a Blood Test Based on Machine Learning. Applied Sciences. Applied Science 2020, 10(23), 8404; https://doi.org/10.3390/app10238404 (This article belongs to the Special Issue Applied Artificial Intelligent). (SJR Q1).

Other contributions made during the development of this thesis were the following:

1. A Framework for Developing Smart Mobile Healthcare Companions for Diabetes Patients at EAI MobiHealth 2020 - 9th EAI International Conference on Wireless Mobile Communication and Healthcare (Accepted).

2. Gamarra, M.; Meriño, I.; Calabria, J. C.; Gutierrez, O.; Barrios, M.; Leal, N.; Wightman, P. Privacy perception in location-based services for mobile devices in the university community of the north coast of Colombia. Ingeniería y Universidad, 23, 1, 2019.

3. As Speaker in Décima Conferencia Iberoamericana de Complejidad, Informática y Cibernética: CICIC 2020, Orlando, Florida, EE.UU. 10 al 13 de Marzo de 2020

4. As Speaker in 2nd Encuentro de Maestrías y Doctorados "Uninorte Investiga" (Barranquilla) at Universidad del Norte, Agosto 2018.

5. As Speaker in Congreso International Interdisciplinariedad y Desarrollo (CIID 2019), Barranquilla, Nov 21-23 de 2019.

6. As Researcher in a fellowship at Group with Strategic Focus in Intelligent Systems in School of Engineering and Science Tecnológico de Monterrey, Campus Monterrey (Mexico) under the supervision of Hugo Terashima Marín (March-Jun, 2019).

7. As Speaker in Research Day at Universidad del Norte(Barranquilla), Octuber 2020.

1.7 | Overview of this thesis

The remaining of this thesis is organized as follows:

Chapter 2 presents the background to understand the main concepts of the thesis, such as the definition of the metabolic syndrome, the criteria for its diagnosis, and its relationship with other diseases. It also shows a metabolic syndrome study carried out by the Universidad del Norte in Barranquilla's population. Moreover, the theories of this framework of machine learning techniques to predict as the performance indicators to evaluate the model and validation techniques.
Chapter 3 presents several data mining methodology and propose a new methodology called RAMAD to predict diseases using machine learning techniques that documents all the phases thoroughly for further improvement of the resulting models.
Chapter 4 shows the results of the literature review process of the search criteria includes the topic of data mining and machine learning techniques applied to predict the metabolic syndrome without using biochemical variables by the scientific community has been developed and published.
Chapter 5 presents the hypothesis of the diagnosis of metabolic syndrome without performing a blood test and compares decision tree, principal components logistic regression, artificial neural networks model with the proposed model using performance indicators.
Chapter 6 presents the MetS segmentation model obtaining ten types of MetS with the HMS criteria and a framework to diagnose each of them without doing a blood test using artificial neural networks Random Subsampling Bosteed Tree comparing each general MetS with performance indicators.

Chapter 7 proposes a SCRUM Thinking framework for creating mHealth applications that contribute to the patient's improvement. Software developers can use this framework to design and implement applications with different software components, some of which could include the prediction algorithms presented in this thesis.

2

Background

This chapter presents the concept of Metabolic Syndrome and a brief study of dataset of this thesis, as well as the theories of the techniques of this framework of machine learning to diagnose the Metabolic syndrome.

2.1 | Metabolic Syndrome

This section presents the base to understand the concept of metabolic syndrome, the prevalence rate worldwide and national, its importance in the medical community, and a brief study of the dataset used in this thesis.

2.1.1 | Definition

Metabolic syndrome is not a worldwide recognized cause of death. Still, it is a trigger that increases the chances of multi-systemically failures, progressively affecting the people affected by it. Such effect creates a pattern of metabolic abnormalities that reflect in the factors associated with the increase in mortality due to diabetes mellitus or coronary heart disease (CHD) (Cornier et al. (2008); Kaur (2014)), which are noncommunicable diseases, and are the leading causes of morbidity and mortality worldwide(WHO (2013)). Patients who have four out of five significant variables have a 3.7 times higher risk of experiencing cardiac events and 24.5 times more risk of being diagnosed with type 2 diabetes (Navarro and Vargas (2008)).

The concept of metabolic syndrome is relatively new, Dr. Gerald Reaven in 1988, presented in the American Diabetic Association an article entitled "the role of insulin resistance in human disease" where it shows the existence of a pattern in a group of signs or symptoms related to the endocrine system which called syndrome X (unknown).

These signs are: resistance to stimulated insulin intake, glucose intolerance, hyperinsulinemia, increased triglyceride levels, decreased HDL cholesterol and hypertension. Reaven concludes that insulin resistance plays a critical role in developing coronary heart disease (Reaven (1993)). The scientific community began to debate their findings, especially cardiologists who began to call this group of symptoms as Reaven syndrome. However, a risk factor was missing, which is the accumulation of abdominal fat (android obesity) proposed by (Kaplan (1989)) and later (Zimmet (1992)) followed by the central perimeter as a factor and also the name change from syndrome X to metabolic syndrome. A few years later, the metabolic syndrome was defined by the World Health Organization (WHO) in its 1st part of the Diabetes report (Alberti and Zimmet (1998)) as a group of clinical diagnostic criteria where it is presented type 2 diabetes or impaired glucose tolerance, plus two factors of the following: hypertension, hyperlipidemy, obesity, and traces of protein in the urine (microalbuminury) as shown in more detail in Table 2.1.

Table 2.1: WHO criteria for diagnosing metabolic syndrome

Variables (Risk Factors)	Decision threshold
Insulin resistance	High levels of insulin or DMT2
Two more risk factors	
Waist/Hip Ratio (WHR)	>0,9 cm men >0,85 women
Triglycerides	>150 mg/dL
HDL Cholesterol	<35mg/dL men
	<39mg/dL women
Systole or Diastole Blood Pressure	>140/90 mmHg
Microalbuminuria	>20pg/min

The WHO also mentions other symptoms such as: hyperuricemia, coagulation disorders, PAI-1 increase, among others. But he recognizes that these are not necessary for the diagnosis of the syndrome. The European Insulin Resistance Consortium publicized that the presence of microalbuminuria was not a requirement to diagnose metabolic syndrome (Balkau and Charles (1999)).

To diagnostic metabolic syndrome also have created several criteria in the history as shown Table 2.2 such as the Adult Treatment Panel of the National Program for Cholesterol Education (ATP III)(Bartlett (2001)), European Group for the Study of Insulin Resistance (EGIR) (Balkau and Charles (1999)) and the International Diabetes Federation (IDF) (Alberti et al. (2006)), which have some small differences in levels some variable. since 2009, experts have a consensus Harmonized diagnostic or Joint Interim Statement

(HMS) to unify the diagnosis of the metabolic syndrome (Alberti et al. (2009)) supported by d the NHLBI (National Institute of the heart, lung and blood) from and the early 1970s, they have collaborated with researchers to develop clinical practice guidelines that focus on the management of cardiovascular disease risk factors (Goff et al. (2014)), in order to unify the diagnosis of metabolic syndrome despite the fact that the ATP III guideline is the most widely used. However, at present the IDF and HMS criteria for the diagnosis of metabolic syndrome at the population level are being used in studies of this condition (Navarro et al. (2013)) which requires central obesity and any two of the four risk factors.

In the current, the Metabolic Syndrome (MetS) is coded E88.81 according to the International Classification of Diseases, 10th Edition (ICD- 10, 2020 version) and is a group of alterations in metabolism and includes dyslipidemia (abnormal concentrations of lipids in the blood: increased triglycerides and decreased HDL cholesterol), hypertension, hyperglycemia, and obesity (Chobanian et al. (2003)). The insidious increase in the elements of MetS, obesity, insulin resistance (IR), and dyslipidemia are responsible for the current global epidemic of type 2 diabetes (Navarro and Vargas (2008)). Other authors relate MetS with the occurrence of cancers and chronic kidney disease (Chen et al. (2004); Esposito et al. (2012)). So, the MetS is relationed with some diseases as shown Figure 2.1 which can be delayed if MetS is detected in order to start treatment to reduce risk factors such as obesity or diet.

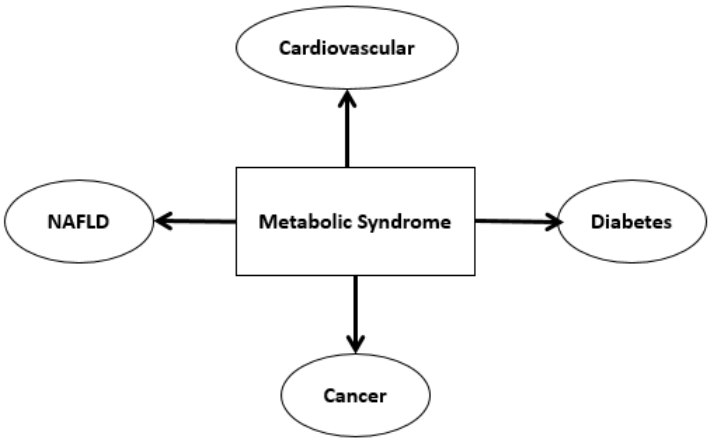

Figure 2.1: Diseases associated to the metabolic syndrome.

Therefore, it is very important to carry out campaigns and create public policies for the prevention of metabolic syndrome and the first steps are studies of the prevalence of metabolic syndrome in the population.

Table 2.2: Criteria of several organizations for diagnosing metabolic syndrome.

Organization	Risk Factors					
	Obesity WC(cm)	TG (mg/dL)	FPG/IGT (mg/dL)	HDL-C* (mg/dL)	BP* (mmHg)	Diagnostic Criteria
EGIR (1999)	M: =>94 F:=>80	>178	>110	<39	SBP>=140 DBP>=90	3 or more Risk Factors
ATP III (2001)	M:=>102 F:=>80	>150*	>110	M:<40 F:<50	SBP:>=130 DBP:>=85	3 or more Risk Factors
AHA (2005)	M:=>102 F:=>88	>150*	>100*	M:<40 F:<50	SBP:>=130 DBP:>=85	3 or more Risk Factors
IDF (2006)	Ethnic specific M:=>90 F:=>80	>150*	>100*	M:<40 F:<50	SBP:>=130 DBP:>=85	Obesity +2 or more Risk Factors
HSM (2009)	Population and Country specific M:=>90 F:=>80	>150*	>100*	M:<40 F:<50	SBP:>=130 DBP:>=85	3 or more Risk Factors

*or treatment; F:Females; M:Males; WC: Waist Circumference; TG: Triglycerides; HDL-C: High-density lipoprotein Cholesterol; FPG: Fasting plasma glucose; IGT: Impaired glucose tolerance; SBP: Systolic Blood Pressure; DBP: Diastolic Blood Pressure

2.1.2 | Prevalence

The available evidence indicates that in most countries, between 20% and 30% of the adult population can be characterized as having metabolic syndrome. In some populations or segments of the population, the prevalence rate is even higher(Grundy (2006, 2008)). The prevalence rate of the metabolic syndrome in countries such as the United States has increased; three studies have yielded the following results: 23.7% in 2002, 34.2% in 2006 and nearly 35% of all U.S. adults were estimated to have the metabolic syndrome in 2011–2012 (Ford et al. (2002); Mozumdar and Liguori (2011)) being this last period estimation of 50% diagnosed MetS in adults older than 60 years of age(Aguilar et al. (2015)).

The prevalence in some Latin American countries are high. For example, Mexico has a 41% (95% CI 0.34 - 0.47) prevalence in adults (Gutiérrez-Solis, Datta Banik (2018)). In Colombia, there are several studies on the prevalence of the syndrome, that focused on specific populations.

For example, a study(Lombo et al. (2006)) with 550 medical records was performed to determine the prevalence of metabolic syndrome as defined by ATP III criteria in comparison with the AHA definition, in patients from the Hypertension Clinic in the Bogotá Santa Fé Foundation University Hospital resulted as shown Figure 2.2, according to the ATP III criteria, the syndrome's prevalence was 27.3% (19.29% males and 30.05% females), while according to the AHA criteria, it was 75.9% (77.9% males and 75.25% females).

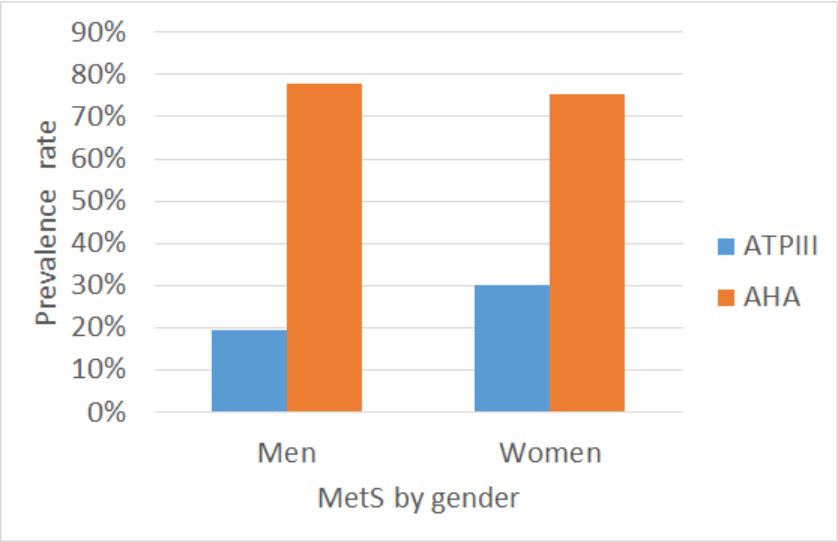

Figure 2.2: Prevalence rate of MetS of by gender in patients hypertension

Other study in Bogota was performed(Lombo et al. (2007)) with 249 clinical histories to determine the prevalence of the metabolic syndrome as defined by the ATPIII compared to the AHA definition in the patients of a third level care institution, diabetes clinic resulting as shown Figure 2.3 the prevalence of the metabolic syndrome according to the ATPIII criteria, was 72, 69% (men 63,83%, women 78,06%), while using the AHA criteria it was 96,77% (men 95,74%, women 96,77%). Morever, 100% of obese and diabetic patients have metabolic syndrome according to the AHA criteria.

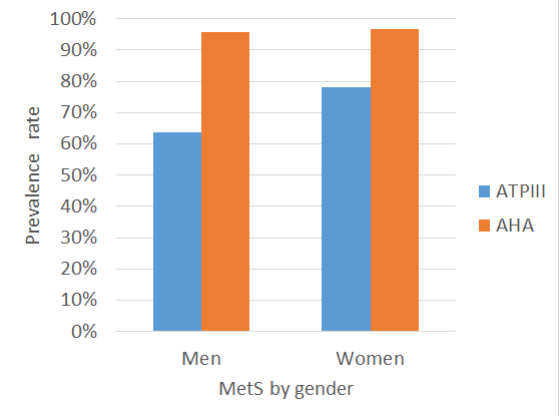

Figure 2.3: Prevalence rate of MetS of by gender in patients diabetes.

Therefore, due to the prevalence of metabolic syndrome found in the population of Bogota and to level worldwide, the Universidad del Norte decided to perform several studies of metabolic syndrome that that explained to continuation.

In the Coast Atlantic of Colombia, a brief study of 62 people in a poor area of Barranquilla, Colombia, found that subjects with at least two risk factors showed a very high prevalence level 74.2% (men 60%, women 78,70%) of MetS based on the ATP III criteria (Navarro and Vargas (2008)) as shown Figure 2.4.

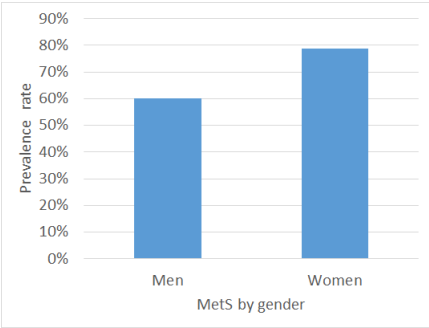

Figure 2.4: Prevalence rate of MetS by gender with at least two risk factors.

A study in other city of the Atlantic of Colombia was performed in Soledad to determine coronary risk in adults with metabolic syndrome. A survey of cardiovascular risks was applied to 99 adults to make measuring of weight, height, waist circumference and blood pressure were taken, as well as biochemical tests for blood glucose, total cholesterol, HDL cholesterol and triglycerides, in order to determine the metabolic syndrome prevalence. Resulting as shown Figure 2.5 a prevalence rate of 49.5% metabolic syndrome according to the IDF, 41.4% according to the AHA, and 20.2% according to the ATP III.

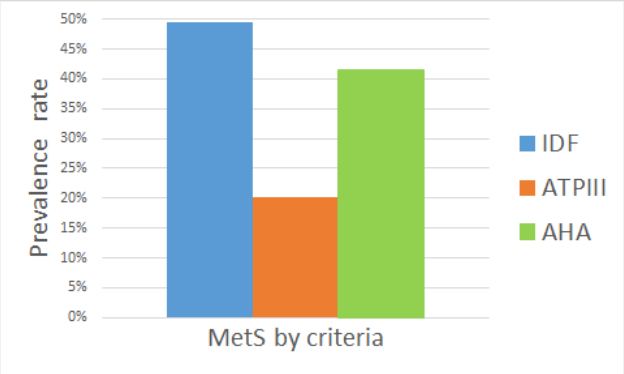

Figure 2.5: Prevalence rate of MetS by criteria.

In addition, the study applied the Framinghan score(Navarro and Vargas (2012)) to evaluate coronary risk found as shown Figure 2.6 some percentage of coronary risk of 8,2% to subjects with metabolic syndrome according to the IDF, 10% according to the AHA, and 7,3% according to the ATP III.

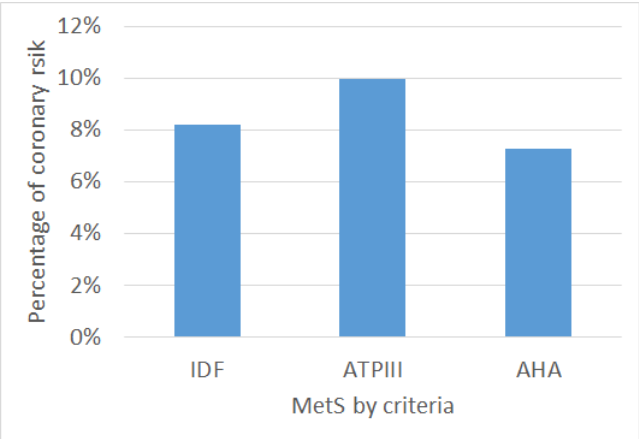

Figure 2.6: Percentage of coronary risk in rank 10% to 20% of subject with MetS by criteria.

In the 2012 year, the Universidad del Norte performed as an integrated research strategy for the study and intervention of the metabolic syndrome in Barranquilla, Colombia, using the following rules:

List of inclusion rules

- Age of 20 years or over.
- The subject can understand the instructions explained by the researchers
- The subject can sign an informed consent
- The subject resides permanently in the area.

List of exclusion rules

- Are you pregnant?
- Are you bedridden?

The study began with a survey of 1,478 adult subjects 20 years old or older randomly selected in 10 city neighborhoods and distributed proportionally according to the neighborhood, and residence block. The survey consisted of several questions divided into several sections. The most relevant sections are obesity history, anthropometric measurements, and biochemical blood measurements such as lipid profile (cholesterol, triglycerides), fasting plasma glucose. The study determined metabolic syndrome and associated factors using the laboratory results plus the surveys' data and the weight measurements, size, and abdominal perimeter. From the sample of 1,478, only 615 (41.61%) decided to continue with the biochemical measurements. So, the study was analyzed 615 records.

The research was carried out under the Good Clinical Practices (GCP) guide and the International Conference on Harmonization (ICH). Therefore, respect for the dignity and the protection of the rights and well-being of people prevailed. The study included protecting the individuals' privacy and autonomy and the decision not to participate in the survey. It is important to note that there was no risk of the participant suffering any damage due to the study. The data that support the findings of this study are available from the corresponding author, upon reasonable request.

This research ensures compliance with the guidelines for the protection of research subjects. Participants received a letter informing them about the project and their rights as participants. The research was approved by the Ethics Committee of the Universidad

del Norte in act 87 in September 2012 and complies with the national guidelines (Resolution No. 8430 of the Ministry of Health of Colombia) and international guidelines (the Declaration of Helsinki) related to the participants' informed consent.

The researchers excluded the following group of people: those who did not give their consent, pregnant women, and those who have suffered a physical or mental incapacitating illness, treatment with steroids, and carriers of decompensated thyroid pathology. The research was carried out according to the Good Clinical Practices (GCP) guide and the International Conference on Harmonization (ICH); therefore, respect for the dignity and the protection of the rights and well-being of people prevailed.

The study included the protection of the individuals' privacy and their autonomy and decision not to participate in the survey. It should be noted that during the investigation, there was no risk of the participant suffering any damage as a result of the study.

2.1.2.1 | Physical examination and blood tests

The respondents arrived at the University del Norte's hospital, where healthcare professionals performed the scheduled clinical examination, executed by a doctor and a nurse. They measured Blood pressure and took two doses with an interval of 5 minutes, averaging the two measurements. Nurses measured stature and weight, without shoes and with the least amount of clothes possible. They also measured waist and hip circumferences.

To collect the information, a structured survey was used, designed by the research group. Four interviewers with training in technical areas of health activity were trained, who had the support of a field supervisor, responsible for reviewing the surveys, in order to detect and correct possible inconsistencies in the formats.

The survey consisted of 190 questions divided into several sections; the most relevant sections are tobacco consumption, physical activity, family history, obesity, anthropometric measurements, and biochemical blood measurements.

From the sample of 1,478, only 615 (41.61%) decided to continue with the biochemical measurements were taken to determine: lipid profile (cholesterol, triglycerides), fasting blood glucose, and serum insulin.

The respondents arrived at the University del Norte's hospital, where healthcare professionals performed the scheduled clinical examination, executed by a doctor and a nurse. They measured Blood pressure, and took two doses with an interval of 5 minutes, averaging the two measurements. Nurses measured stature and weight, without shoes and with the least amount of clothes possible.

Height was measured with a height rod and weight and body fat with a Tanita Ironman electronic scale, with a precision of 5 grams, without shoes.

For the waist circumference, a measuring tape graduated in centimeters was used, with the subject standing and the arms in anatomical position, measured at the midpoint between the anterior superior iliac spine and the inferior coastal margin.

Blood pressure was determined with a previously calibrated mercury manometer, and two measurements were taken with an interval of 5 minutes, with the subject seated with a back, on the right arm, averaging the two figures, according to the recommendations of the Joint National Committee VII.

2.1.2.2 | Data Description

The analysis of the data from the patients who were diagnosed based on the HMS criteria were divided into two populations to explain the prevalence by gender, consisting of a total of 348 women and 267 men between 20 and 96 years of age. Besides that, from the data, 262 presented a diagnosis of MetS, and 353 were used as control according to HMS.

The biochemical variables of triglycerides, fasting plasma glucose, HDL-C were obtained through blood tests obtaining results described in three groups: MetS, Non-MetS, and total as shown in Table 2.3.

Table 2.3: Statistic description of the biochemical variables

Variables*	**MetS m(SD)**	**No MetS m(SD)**	**Total m(SD)**	**p**
TG	216.94(112.8)	121.84(63.05)	160.81(98.67)	<0.001
GL	97.33(38.82)	84(19.56)	89.47(29.74)	<0.001
HDL-C[W]	38(8.22)	46.97(13.54)	43.39(12.49)	<0.001
HDL-C[M]	36.11(11.1)	43.32(11.63)	40.27(11.93)	<0.001

TG: Triglyceride; GL: Fasting Blood Glucose; M: Men; W: Women; Average(m); Standard deviation (SD); *(mmol/L)

The biochemical variables present between the MetS' group and the Non-MetS group have different statistical significance. These variables, together with the waist circumference and systolic and diastolic blood pressure, are necessary to diagnose the MetS, and the study found the total prevalence rate was 42.60% divided into 44.94% for men and 40.8% for women as shown Figure 2.7.

This study of the metabolic syndrome of 615 subjects with the anthropometric and clinical variables was entered into a database that will be analyzed and processed in a framework of machine learning to diagnose the metabolic syndrome, which will be detailed in the next chapters.

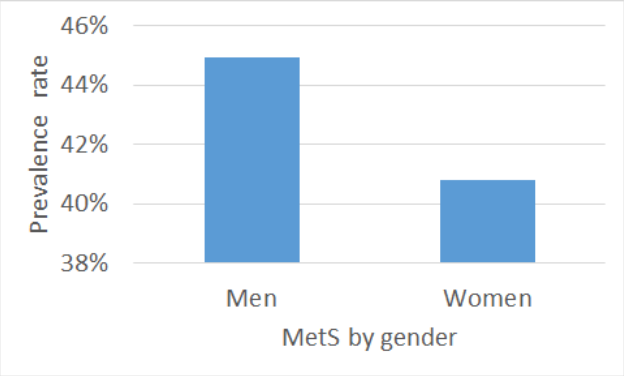

Figure 2.7: Prevalence rate of MetS by HMS criteria.

2.2 | Machine learning techniques

This section presents the techniques and concepts of machine learning used in this thesis. In particular, the concepts of decision tree, logistic regression, artificial neural networks and ensamble learning technique. Combinations analysis and minimization of Boolean equation and validation techniques of models are also described.

2.2.1 | Decision Tree

The decision tree classifies data into a set and determines the values that a variable will take from a data entry. It is a process where decisions are made in sequence descending through the tree.

2.2.2 | Principal Component Logistic Regression

2.2.2.1 | Principal Component Analysis

The dataset is transformed into new variables, which generate the original vectors through a linear combination as shown in the Equation (2.1) and are selected for their variance ($\sigma \geq 1$).

$$PC_n = C_0 * X_0 + C_1 * X_1 + ...C_n * X_n \tag{2.1}$$

2.2.2.2 | Logistic Regression

These new variables are orthogonal and apply to the technique of logistic regression that the result of nominal qualitative type depends, or not, on other predictor variables, that is, PC_n. The nature of the predictor variables can be dichotomous and quantitative. The Equation (2.2) represents the logistic regression where P is the probability of occurrence of a true positive in Equation (2.3) and β_n are the regression coefficients through the Wald statistic.

$$\ln \frac{P}{1-P} = \beta_0 + \beta_1 * PC_1 + ...\beta_n * PC_n \tag{2.2}$$

$$P = \frac{e^{\beta_0 + \beta_1 * PC_1 + ...\beta_n * PC_n}}{1 + e^{\beta_0 + \beta_1 * PC_1 + ...\beta_n * PC_n}} \tag{2.3}$$

2.2.3 | Multilayer Perceptron Artificial Neural Network

The Multilayer Perceptron (MLP) Artificial Neural Network (ANN) are used to predict.the depend variable. These ANN should be trained before being used to predict the output variable value, that is, the dependent variable. Each ANN is formed by neurons whose elements are a set of inputs that can come from other neurons or the outside, as shown in Figure 2.8 the basic structure of an ANN.

Each structure of ANN should be initialized according to the propagation rule to the starting and each node has synaptic weights, which are the degree of communication between neurons, as shown the Eqs. 2.4 and 2.5. Then,the data used to train the network is introduced into the network after the propagation algorithm is employed to obtain the final parameters in the network. In practice, the algorithm is divided into two parts:

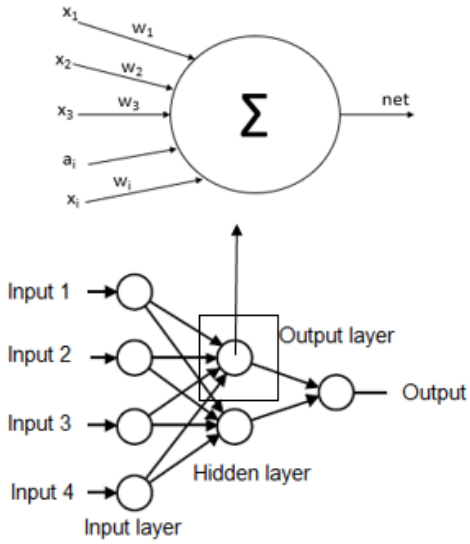

Figure 2.8: Basic structure of the artificial neural network

network training and network testing. The steps of propagation algorithm are described as follows(Kumar (2012)):

$$net^k = \sum_{i=1}^{n} (\omega_i^k x_i^k - \alpha_i^k) \tag{2.4}$$

$$y^k = \theta(net^k) \tag{2.5}$$

Where x_i^k are inputs, ω_i^k are synaptic weights, α are bias in the input layer, k is the iteration and n is the number of inputs resulting in a net output, that is determined by a activation function $\theta(net^k)$ with output y^k (Bishop (2006); Friedman, J., Hastie, T., & Tibshirani (2009)).

This information flows in one direction only from the inputs to the hidden layer and after to the output layer, that is, the information that comes from different activation function neurons, which is responsible for determining the current state and finally converges all the data to the output (Ivanović et al. (2016)).

Each ANN has several hidden neurons that have functions, such as the hyperbolic tangent sigmoid function and an output layer with a neuron. The neuron has a function that can be a log-sigmoid function(Bishop (2006); Witten and Frank (2005)).

It should be noted that there are no hard and fast rules for the number of hidden neurons. These hidden neurons can be calculated or found empirically and are highly

dependent on the problem and the dataset(Andrea and Kalayeh (1991)). However, I used the methodology mentioned by (Boger and Guterman (1997); Karsoliya (2012); Panchal and Panchal (2014)) and described in Eq. 2.6, where the number of hidden neurons (NHN) can be 2/3 of the input variables plus an output variable.

$$NHN = \frac{2(Input\ variables)}{3} + Output\ variables \tag{2.6}$$

In the Eq. 2.6, are used to estimate the number of hidden neurons to contribute to research in the area of machine learning for the diagnosis of MetS without using biochemical variables and in a way, describe every detail of the process for experimentation by other researchers can continue investigating these models as well as Chen(Chen et al. (2014)) that used other equation to calculate the hidden neurons.

2.2.4 | Random Undersampling Boosted Tree

Another machine learning technique used to diagnose the MetS types was the ensemble Random undersampling Boosted tree (RusBoost) because the data from the MetS study is imbalanced(Mounce et al. (2017)). This technique improves the performance indicators of models using imbalanced data by applying a random undersampling technique. The technique randomly removes samples from the majority class(Seiffert et al. (2008)), as shown in the algorithm detailed in Appendix A with the configuration showed in Table 2.4.

Table 2.4: RusBoost Configuration

Learned Type	Decision tree
Maximum number of splits	20
Number of learners	30
Learning rate	0,1

2.3 | Model Validation

Any model that is propagated must be validated and, thanks to mathematical sciences, different model validation techniques have been developed, which are explained below:

2.3.1 | Hold Out

The original dataset is split into two different datasets labeled as a training and a testing dataset. This can be a 60/40 or 70/30 or 80/20 split with a random distribution for then to obtain the performance indicators of the model.

2.3.2 | Random Subsampling Validation

For the validation of the model, it was used random subsampling or Monte Carlo cross-validation on multiple data that are randomly chosen from the dataset and combined to form a new dataset, that is, multiple hold outs. The remaining data forms the training 70% and testing 30% of the dataset. The test data predictions give a realistic estimate of the external validation data predictions because it is asymptotically consistent. This approach results in more pessimistic predictions of the test data compared to cross-validation (Berrar (2019); J. Shao (2005); Liang and Y.-Z. (2001); Park and Kim (2012)).

2.4 | Performance Indicators and Model Assessment

The indicators to evaluate the capacity for discrimination that have the models, such as Sensitivity (SS), Specificity (SP), False Negative Rate (FNR), False Positive Rate (FPR), Accuracy, AROC(Perveen et al. (2019); Witten and Frank (2005)). The TP, TN, FP, and FN values represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

$$Sensitivity(SS) = \frac{TP}{TP + FN} \tag{2.7}$$

$$Specificity(SP) = \frac{TN}{TN + FP} \tag{2.8}$$

$$False\ Positive\ Rate(FPR) = 1 - SP \tag{2.9}$$

$$False\ Negative\ Rate(FNR) = 1 - SS \tag{2.10}$$

$$Positive\ Predictive\ Value(PPV) = \frac{TP}{TP + FP} \tag{2.11}$$

$$Negative\ Predictive\ Value(NPV) = \frac{TN}{TN + FN} \tag{2.12}$$

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.13}$$

$$Area\ Receiver\ Operating\ Characteristic(AROC) = \int (SS)(1 - SP) \tag{2.14}$$

The classifiers used to diagnose diseases prioritize the rate of type 1 and 2 errors known as False Negative Rate (FNR) and False Positive Rate (FPR), which are the complement of SS and SP, respectively. The type 2 error (false negative) is harmful because a patient with the disease can be diagnosed as a patient without the illness, affecting their health since they did not start adequate treatment.

On the other hand, it should be noted that all these models were evaluated using the AROC and the Hosmer and Lemeshow criterion(D. W. Hosmer and Lemeshow (2004)), which is shown in Table 2.5.

Table 2.5: Assessment rules of AROC.

AROC	**Discrimination Ability**
AROC=0.5	No discrimination
0.5 < AROC < 0.7	Regular
$0.7 \leq$ AROC < 0.8	Acceptable
$0.8 \leq$ AROC < 0.9	Excellent
AROC ≥ 0.9	Outstanding

2.5 | Summary

In summary, worldwide is recognized that metabolic syndrome is the trigger to increase the chances of developing heart disease and diabetes mellitus(Cornier et al. (2008); Tagle-Luzárraga et al. (2007)). Therefore, it is imperative to develop tools to diagnose metabolic syndrome early to avoid or reduce their consequences. Thanks to computer science it is possible to develop a framework of machine learning to determine the metabolic syndrome early using the techniques explained as the validation techniques, and the performance indicators.

3

A Novel Data Mining Process Methodology

Many researches carried out to propose a machine learning model to diagnose a disease do not explicitly use a data mining methodology. So, it is very common to find article of the same topic with different parameters to assess the models used to diagnose disease.

This chapter presents several datamining process methodology and propose a new methodology to diagnose diseases using machine learning techniques that documents all the stages thoroughly for further improvement of the resulting models.

3.1 | Introduction

In a society where everything is recorded in large databases and the health sector, it is not the exception since the database stores the data of the medical record that stores all the relevant information about the patient's health along with other data to be analyzed and obtain pertinent information to improve the conditions of patients and institutions. To acquire this information, the resarcher using a datamining process methodology. So, it is very important to know the different datamining process methodology to choose the one that adapts the most to homogenize the different models of machine learning that can be found in a literature review or propose a new methodology.

3.2 | Datamining Process Methodologies

In datamining, there are several methodologies such as Knowledge Discovery in Databases (KDD), Cross Industry Standard Process for Data Mining (CRISP-DM), and Sample Explore Modify Model Assess (SEMMA)(Calabria and Bonilla (2014); Stirrup and Ramos

(2017); Wirth (2000)). However, these are not explicitly used in the analyzed articles of the literature review of the thesis because usually authors look for a technique goal, and they do not have a clear business goal as expected (Calabria and Bonilla (2014); Fayyad et al. (1996); Shearer (2000)).

3.2.1 | KDD Methodology

KDD is a non-trivial process of efficiently and consistently identifying potentially useful and previously unknown patterns in databases. This process is made up of several stages in way of general as shown in figure 3.1.

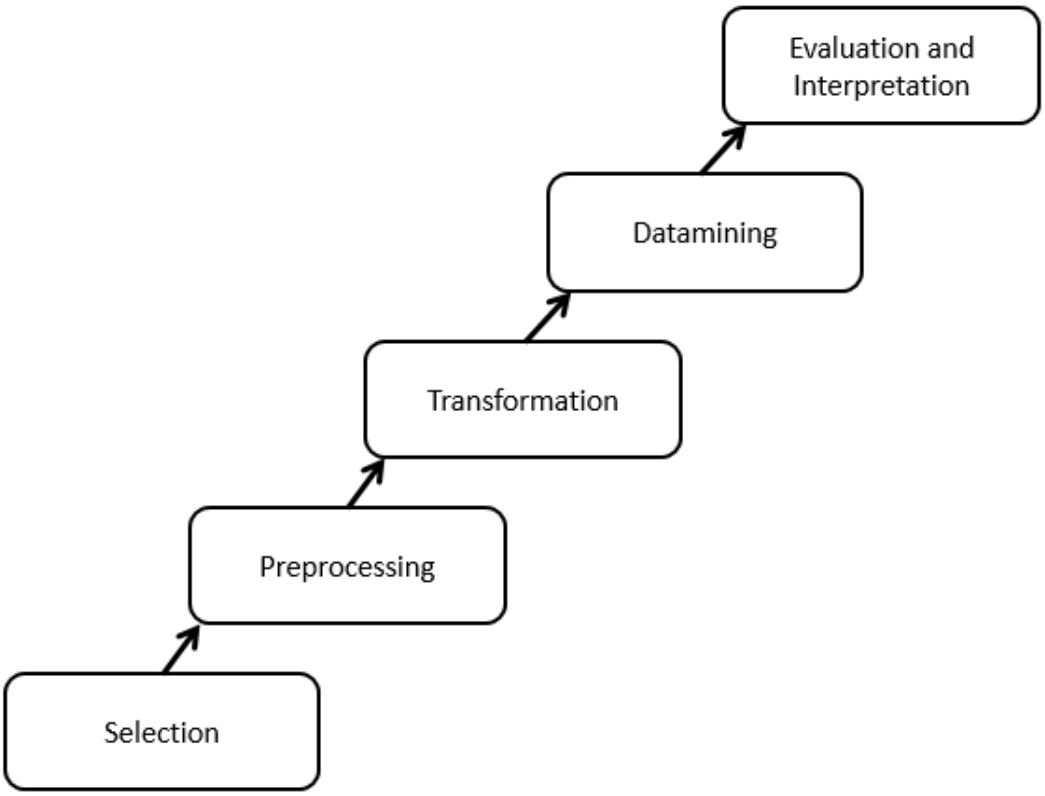

Figure 3.1: Stage of the KDD Methodology(Azevedo, A. and Santos (2008)).

- **Selection:** This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
- **Preprocessing:** This stage consists on the target data cleaning and pre processing in order to obtain consistent data.

- **Transformation:** This stage consists on the transformation of the data using dimensionality reduction or transformation methods.
- **Data mining:** In the data mining stage, the task to be carried out is decided (classification, cluster, prediction) and the method to be used is chosen.
- **Evaluation and Interpretation:** In the evaluation and interpretation stage, the patterns are evaluated and analyzed by the experts and if necessary, it returns to the previous stages for a new iteration.(Azevedo, A. and Santos (2008); Calabria and Bonilla (2014))

This KDD process demands that each stage be carried out to obtain useful knowledge, the datamining stage being so important. However, many researches carried out to propose a machine learning model to predict a disease do not explicitly use a data mining methodology. So, it is very common to find article of the same topic with different parameters to assess the models used to diagnose disease.

3.2.2 | CRISP-DM Methodology

CRISP-DM is a free distribution reference data mining methodology, the most widely used in the development of multi-stage composite data mining projects as can be seen in figure 3.2.

- **Business understanding:** This stage includes determining business objectives, evaluating the current situation, setting data mining goals, and developing a project plan.
- **Data understanding:** Here, it must understand the sources of data, in order to become familiar with them. This step can include initial data collection, data description, data exploration, such as viewing summary statistics (which includes viewing the categorical variables) can occur at the end of this stage. Models such as cluster analysis can also be applied at this stage, with the intention of identifying patterns in the data.
- **Data preparation:** The idea is to build the final dataset, to which the selected modeling tools will be applied in the next stage. Their tasks include: selecting and cleaning the data, building and integrating new data, if necessary, and finally, applying the necessary conversions so that the data complies with the required format.

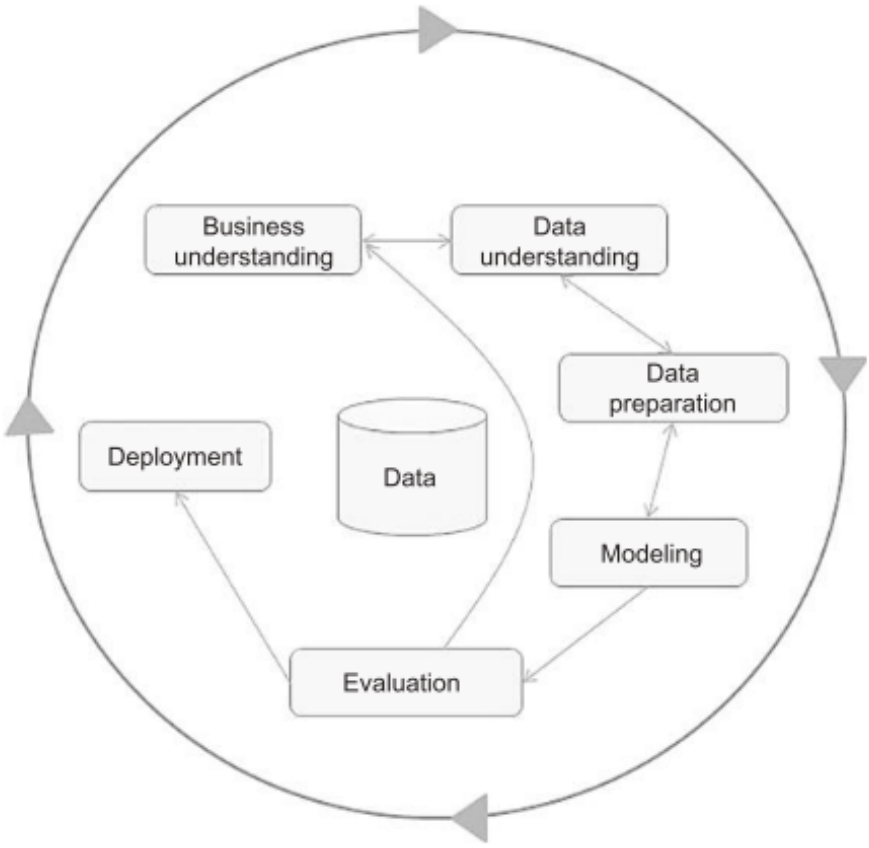

Figure 3.2: Stage of the CRISP methodology (Kotu and Deshpande (2014)).

- **Modeling:** The model is built from the selected techniques and the data previously prepared; Within this stage, the following tasks are carried out: select modeling techniques, generate the test design, build and interpret the model. By performing the modeling stage, the appropriate datamining technique can be selected to obtain the predictive models through the training processes.

- **Evaluation:** The model is evaluated taking into account the business objectives, the process is reviewed to determine improvements and the next stage to be carried out is defined taking into account the results obtained in the evaluation.

- **Deployment:** Finally, monitoring and maintenance activities are carried out, the results are organized, and the results are shown generating reports that identify the success or failure of the project. The deployment will allow the organization and storage of the obtained predictive models for later use(Calabria and Bonilla (2014)).

Many research projects that apply data mining techniques to predict diseases do not explicitly use a methodology as CRISP-DM but follow a similar approach to those found in the literature as SEMMA.

3.2.3 | SEMMA Methodology

The SEMMA methodology is the most similar to the ones used in the articles found in this thesis's literature review. SEMMA is the acronym for the five stages shown in Figure 3.3 and is promoted by SAS Institute Inc. (Matignon (2007); Rohanizadeh and Moghadam (2009)). Many articles explore the data and, in some cases, modify it with the creation of new variables to model and evaluate their ability to diagnose. Figure 3.3 shows the SEMMA methodology, and the stages are explained next.

Figure 3.3: Stages of the SEMMA methodology(Calabria and Bonilla (2014)).

- **Sample** This step consists of sampling the data by extracting a dataset big enough to contain meaningful information, and small enough to be processed quickly.
- **Explore** the data by searching for anticipated relationships, unanticipated trends, and anomalies to gain understanding and ideas.
- **Modify** the data by creating, selecting, and transforming the variables to then focus on the model selection process.
- **Model** This step consists of modeling data to find data combinations or patterns that reliably predict the desired outcome.
- **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process (**?**).

3.2.4 | RAMAD Methodology

The methodology proposed in this thesis is different from SEMMA as it proposes an incremental circular approach for continuously checking for new models in the literature. Those models are used to compare with the researchers' data to improve the

models based on the newly found information with each new iteration of the methodology, allowing for the reuse of the model with other data. Figure 3.4 shows the proposed methodology, which consists of the following stages: Review, Analyze, Model, Assess, Document (RAMAD).

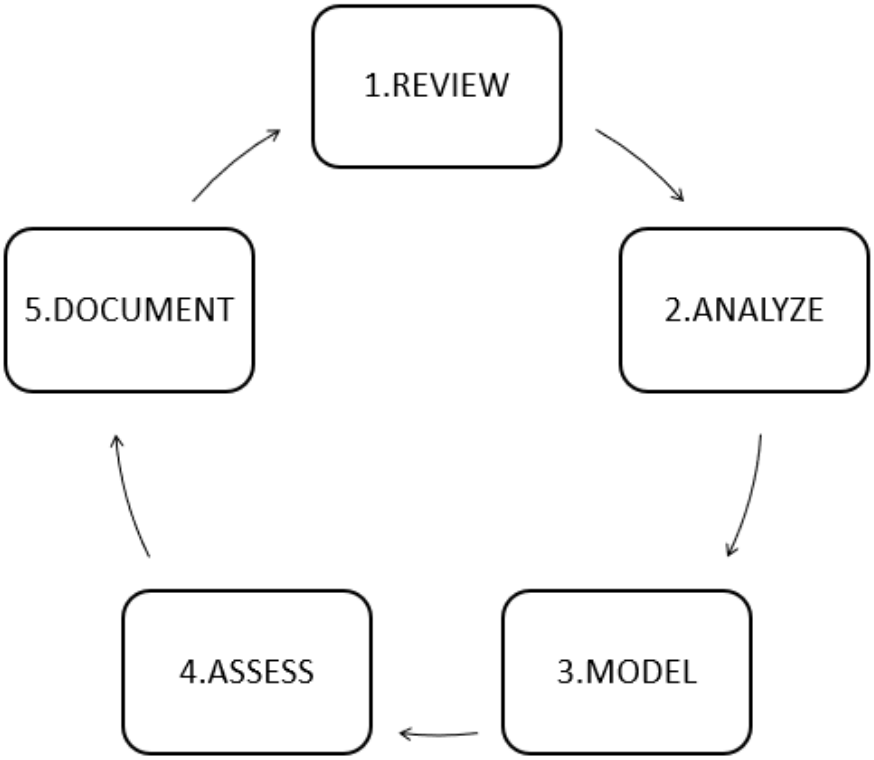

Figure 3.4: Stage of the RAMAD methodology.

1. REVIEW: This stage should start with a research question for the studied disease. Then, the user of the methodology performs a literature review that collects and revises the works proposing predictio (diagnosis) models for the disease. It should also describe how to obtain the data of the target population, which could come from publicly available data or data obtained as part of the project.

2. ANALYZE: It is a data analysis that includes a data description, correlation between data, and other tests.

3. MODEL: It consists of building several classification models and testing them using the obtained data. If it is possible, propose a classifier model to detail which models are better used to predict the disease.

4. ASSESS: In this stage, researchers must assess the proposed models and perform a comparison between them using the selected performance indicators.

5. DOCUMENT: Record the parameters of the models and verify that they are detailed so that the project and experiments can be replicated.

The execution of the methodology can be updated and is encouraged by starting again in the Review stage and adding new models from the literature. This circular approach should ensure that researchers keep updating original models, and with appropriate documentation, new researchers could contribute to an improved diagnosis model.

This methodology called RAMAD does not intend to replace the KDD, CRISP-DM and SEMMA methodologies since it is only a methodology that proposes to homogenize the experiments based on machine learning of the researches on some disease so that each new investigation can contribute a little more to the search for a general model on the prediction of disease

3.3 | Summary

The proposed methodology looks into creating a research tool for research projects solving health-related problems with classification models to predict diseases. Such projects choose their predictor variables based on the relationship between the data and the outcome variable, then compare their results with other authors' results. The methodology helps the research community by building a classification model from the literature's variables and comparing the proposals using data from a particular study on the disease. This approach contributes to the generalization of a classification model that can be implemented on a large scale. Therefore, I expect that researchers can use the RAMAD methodology to get a generalized model for predicting metabolic syndrome or related diseases.

4

Literature Review

The academic community has preform several researches and had cited the metabolic syndrome in more than 24,000 publications registered in PUBMED until May 2020. Of course, the number of publications or citations related to the metabolic syndrome provides an estimate of the importance of the subject as research in the scientific field. Metabolic Syndrome (MetS) is a cluster of risk factors that increase the likelihood of heart disease and diabetes mellitus. Moreover, it is a disease that can trigger other diseases or conditions such as Nonalcoholic Fatty Liver Disease (NFLD)(Boutari et al. (2018)).

Therefore, it is important to carry out studies to identify the prevalence of metabolic syndrome in the population early in order to take measures at the level of public strategies that can save lives.So, it is crucial to get diagnosed with time, that is say, diagnosing early, to take preventive measures, especially for patients in locations without proper access to laboratories and medical consultations because for the diagnosis of metabolic syndrome, 5 variables are required of which 3 are obtained through a blood sample. But, what solution to the problem of predicting the metabolic syndrome early by the scientific community has been developed and published.

4.1 | Literature review process

This chapter shows the literature review process using the proposed approach by (Reyes et al. (2017); Urru (2011); Webster and Watson (2002)) and it is described in the Figure 4.1. The search criteria includes the topic of data mining and machine learning techniques applied to the disease. In this work, the studied disease was metabolic syndrome, and I was interested in models which did not use a blood sample. It was made using four

search engines: DBLP, IEEE XPLORE, ACM and PubMed with a window of 12 years counted retroactively from 2019.

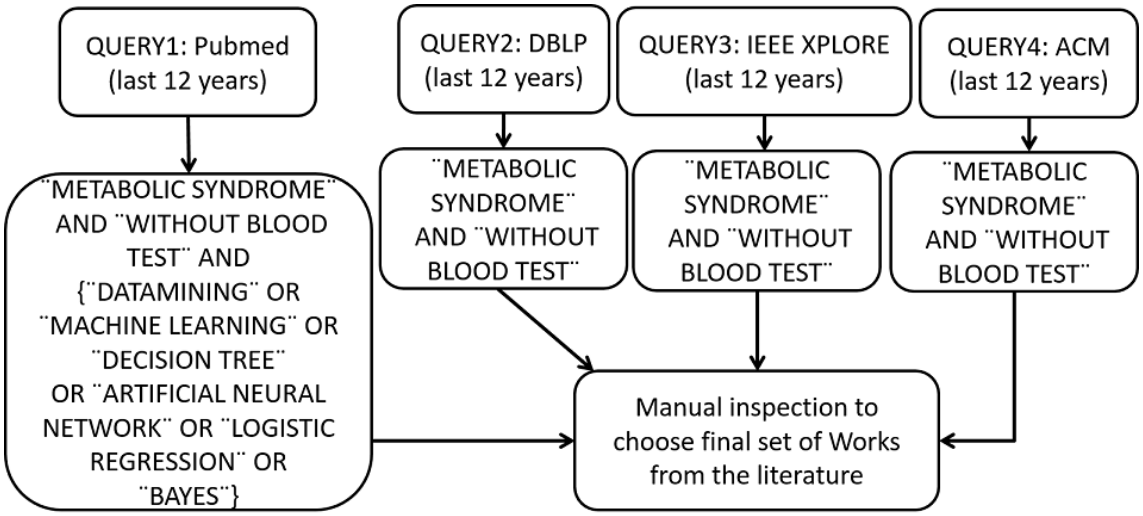

Figure 4.1: Selection criteria for choosing classification models for MetS diagnosis from the literature.

The keywords used in the search process were metabolic syndrome, without a blood test, data mining, and machine learning due to the similarities associated with data mining. To overcome this difference, a search for articles in PubMed (QUERY1) was carried out with keywords "METABOLIC SYNDROME" AND “WITHOUT BLOOD TEST” AND (“DATA MININ” OR “MACHINE LEARNING” OR “DECISION TREE” OR “ARTIFICIAL NEURAL NETWORK” OR “LOGISTIC REGRESSION” OR “BAYES”). Moreover, in DBLP, IEEE XPLORE, and ACM I searched for “METABOLIC SYNDROME” AND “WITHOUT BLOOD TEST” (QUERY2), (QUERY3) and (QUERY4) without mentioning “data mining” and “machine learning” since these keywords would generate too many results. Manual inspection was performed and then filtered by the criteria established as shown in Table 4.1.

4.2 | Results

The results of the queries (QUERY1: 101, QUERY2: 51, QUERY3: 90 and QUERY4: 81) provided a large list of articles and conferences. However, not all the documents had a direct relationship with the metabolic syndrome since DBLP, IEEE XPLORE and ACM delivered articles not only of data mining and machine learning but of articles of computer science at a general level. Then the delivered list was filtered, evaluating

Table 4.1: Criteria established for choosing the articles.

Criteria for the Selection of Articles
Can be obtained the variables through first-level medical attention? Y/N
What diagnostic criteria do authors use? e.g., ATP III, IDF, HMS or other criteria recognized.
What machine learning technique do authors use? e.g., logistic regression, decision tree, artificial neural networks, and others
What validation method do authors use? e.g., hold out, random subsampling and others.
What performance indicators do authors use? e.g., Area under the ROC curve, sensitivity, specificity, positive and negative predicative values and accuracy

its relationship with the syndrome, and also excluded repeated articles (mirror articles). The manual inspection documented in Table 4.1, which reduced the list to a collection of 4 articles, gives the specific requirements to be included in the final list. Each article was evaluated based on the established criteria. These articles used classification models for the diagnosis of metabolic syndrome without doing a blood sample using variables obtained in a medical consultation, as shown on the Table 6.2. The table shows the variables and if the authors used them explicitly in their model, or implicitly.

Kroon et al. (Kroon et al. (2008)) proposed a decision tree model to eliminate the need for blood tests for the diagnosis of the MetS in 50–90% of the cases. The variables used were BMI, WCD, SBP, BPD, and in an implicit way, SEX, and WC due to thresholds of WC and also SBP and DBP. The authors conducted this study with 642 young adults between 17–28 years old. Diagnosis of MetS was defined according to NCEP ATP III with a sample that only had 7.48% of the dataset with MetS (48/642). It is important to remark that subjects with BMI ≥ 35 all had MetS. This paper was the first to use a machine learning model in the problem of diagnosis of the MetS without biochemical variables. However, the authors did not report the model's performance measure variables used in their dataset.

Murguia-Romero et al. (Murguía-Romero et al. (2013)) build several tools to predict MetS in young Mexicans using BMI, WC, Weight, Height, Sex, and other non anthropometrics variables. However, it does not take into account the variables of systolic and diastolic pressure which are risk factors for the diagnosis of MetS. They configured an Artificial Neural Network (ANN) based on Multilayer Perceptron (MLP) with back propagation and trained it with diagnosed patients using the HMS criteria with 70%.

Table 4.2: Variables and technique used by the authors.

Authors	**Kroon**	**Murguia-Romero**	**Chen**	**Kupusinac**
Age			E	E
Sex	I	E	E	E
Weight	I	E	I	I
Height	I	E	I	I
WHR: Waist to Hip ratio			E	
WSR: Waist to Stature ratio				E
BMI: Body mass index	E	E	E	E
HC: Hip circumference			E	
WC: Waist circumference	I	E	E	I
WCD: WC Dichotomous	E			
BPD: Blood Pressure Dichotomous	E			
SBP: Systolic blood pressure	I		E	E
DBP: Diastolic blood pressure	I		E	E
Technique	DT	BPNN	PCLR, BPNN	BPNN

DT:Decision Tree; PCLR: Principal Component Logistic Regression; BPNN: Back Propagattion Neural Networks; E: Explicit use of variable; I: Implicit use of variables

They then tested the model using 30% of the set, using a total of 826 people. They used the Positive Predictive Value (PPV) as a performance indicator which varies from 38.2% to 45.4%. Also, the authors used a particular ANN of 25 hidden neurons using BMI, WC, Weight, Height, and Sex getting an average PPV 38.8 with a standard deviation 12.8.

Chen et al. (Chen et al. (2014)) proposed a neural network to diagnose the metabolic syndrome without using biochemical variables such as blood glucose and cholesterol levels. The authors proposed instead the use of anthropometric variables such as Sex, Age, BMI, WC, HC, WHR, SBP, and DBP, and in an implicit way, they used Weight and Height to measure the occurrence of the metabolic syndrome. They compared the technique of Principal Component Logistic Regression (PCLR) with neural networks for predicting Met with criteria IDF using a dataset of 2074 individuals (male: 1495, female: 579), obtaining improved results in the BPNN.

Hsiung et al. (Hsiung et al. (2015)) through a sample of 154 subjects with a prevalence 40.26% of MetS segmented in four groups depending on the number of factors of the criteria for the metabolic syndrome with criteria ATP III with variables from physical evaluation, lifestyle profile, heart rate variability, and blood analysis. After excluding the invasive blood tests; the results of multivariate logistic regression identified like non-invasive evaluation variables (blood pressure, body mass index and very lower frequency of heart rate variability) that were significant predictors for the risk of suffering from the metabolic syndrome. This work is noteworthy because it relates the characteristics of the variability of the heart rate with the metabolic syndrome as well as the variable Neck Circumference (NC). However, the authors did consider this variable, which can not be obtained with equipment that is in the first level of medical attention. Therefore, this work is excluded from the analysis of the comparisons with the other models.

Kupusinac et al. (Ivanović et al. (2016)) developed an Artificial Neural Network (ANN) to predict MetS that includes non-invasive variables that easily can be obtained by applying low-cost diagnostic methods. The ANN entry vectors are sex, age, BMI, WHR, SBP, DBP, and in an implicit way WC, Weight, Height due to the use of WSR and BMI. The ANN output is dichotomous in true/false for the prediction of the diagnosis of the MetS with criteria IDF. The ANN training, validation, and testing are carried out in the large dataset that includes 2,928 people, and the authors built a Feed-forward ANNs with 1–100 hidden neurons. The solution achieved the highest positive predictive value PPV = 0.8579 and a Negative Predictive Value NPV = 0.8319.

4.3 | Summary

Most of the articles found in the literature review, of machine learning models to diagnose metabolic syndrome without using a blood sample, use a methodology similar to SEMMA due to the similarities of the phases of these articles with the phases of SEMMA in its development. However, they do not mention a methodology to obtain specific machine learning models in their articles. Furthermore, none of the articles tested the existing models in the literature to advance the search for a generalized model with data from multiple populations in order to contribute to developing more robust models.

5

Prediction of Metabolic Syndrome without doing a Blood Test

In the literature review, several machine learning models were found to diagnose metabolic syndrome without doing a blood test with some indicators of performance. However, these indicators were not in all articles to evaluate the models. Moreover, none of the articles worked by contributing to existing models to advance in the search for a generalized model with data from multiple populations. Such a type of approach could contribute to more robust models.

Therefore, this chapter seeks to homogenize the evaluation of the models found in the literature review and uses the RAMAD methodology to perform several experiments using the data of the study of metabolic syndrome described in chapter 2, obtaining results that are compared with the proposed machine learning model and finding the best machine learning technique to diagnose MetS without explaining (Palacios et al. (2017)).

5.1 | Introduction

Machine learning classification models have been used for the diagnosis of several diseases and conditions (Chandna (2014); Smith et al. (1988)). One type of classification model is to study the probability of a diagnosis of Metabolic Syndrome (MetS). MetS is a group of alterations in metabolism that includes dyslipidemia (abnormal concentrations of blood lipids: increased triglycerides and decreased HDL cholesterol), hypertension, hyperglycemia, and obesity (Chobanian et al. (2003)). These are known as metabolic risk factors that increase the likelihood of heart disease or diabetes mellitus (Cornier et al. (2008)) since the syndrome indicates a 5-fold increase in the risk of type 2 diabetes mel-

litus (T2D). Thus, it is often diagnosed as prediabetes (Aschner (2010); Grundy (2007)). Also, the syndrome doubles the risk of developing Cardiovascular Disease (CVD) (Kaur (2014); Tagle-Luzárraga et al. (2007)). Other authors relate MetS with the occurrence of cancers and chronic kidney disease (Chen et al. (2004); Esposito et al. (2012)). Therefore, it is vital to develop mechanisms to achieve the identification of the MetS early in order to avoid or delay its appearance already mentioned by many authors of the scientific community (Aschner (2010); Galassi et al. (2006); Jover et al. (2011)).

The prevalence of the syndrome in countries such as the United States has increased; three studies have yielded the following results: 23.7% in 2002, 34.2% in 2006 and nearly 35% of all U.S. adults were estimated to have the metabolic syndrome in 2011-2012 being this last period an estimation of 50% diagnosed MetS in adults mayor of 60 years of age (Aguilar et al. (2015)).

These studies even showed that the prevalence of MetS is higher among the Mexican-American population. In Mexico has a 41% (95% CI 0.34–0.47) prevalence in adults, and in some Latin American countries (Gutiérrez-Solis, Datta Banik (2018)) such as Colombia, several studies on the prevalence of the syndrome were performed focusing on specific populations. For example, a brief study of 62 people in a poor area of Barranquilla, Colombia, found that subjects with arterial hypertension showed a very high prevalence level (74.2%) of MetS based on the ATP III criteria (Navarro and Vargas (2008, 2012)).

Metabolic Syndrome is diagnosed with a set of risk factors whith some threshold levels in the criteria proposed by several medical associations being recently changed as shown in the Table 5.1. Some associations were used in the articles found in literature review such as the Adult Treatment Panel of the National Cholesterol Education Program (ATP III) (Bartlett (2001)), International Diabetes Federation (IDF) (Alberti et al. (2006)), and the consensus of Harmonized Diagnosis (HMS), to unify the diagnosis of the syndrome (Alberti et al. (2009)).

For diagnosing, MetS must have at least three of the five conditions of the risk factors, except IDF, which requires the central obesity and any two of the four risk factors leftover as shown in Table 5.1 and emulated using the algorithm shown in Appendix B and Appendix C to diagnostic with IDF (Ivanović et al. (2016)) and HMS criteria, respectively.

Table 5.1: Diagnosis criteria.

Risk Factors	**ATP III(2001)**	**IDF(2006) and HMS(2009)**
Central Obesity WC Waist Circumference	Male: WC $\geq$ 102 cm Female: WC $\geq$ 88 cm	Country/ethnic,specific values for WC South American population Male: WC $\geq$ 90 cm Female: WC $\geq$ 80 cm
TG Triglycerides	>150 mg/dL	>150 mg/dL or treatment
FG Fasting Glucose	>100 mg/dL without diabetes	>100 mg/dL or treatment
HDL-C High-density lipoprotein cholesterol	Male:<40 mg/dL Female:<50 mg/dL or treatment	Male:<40 mg/dL Female:<50 mg/dL or treatment
BP, SBP, DBP Blood Pressure Systole and Diastole Blood Pressure	SBP $\geq$ 130 mmHg and DBP $\geq$ 85 mmHg or treatment	SBP $\geq$ 130 mmHg or DBP $\geq$ 85 mmHg or treatment

The diagnosis always relies on five factors, independent from the criteria used. Of the five elements, two (Waist Circumference (WC) and Blood Pressure (BP)), are obtained in a medical consultation. The other three factors (triglycerides, HDL-C, and glycemia) require invasive tests to know their value in the patient's blood. The complete diagnosis implies a second medical visit because it is necessary to wait for the test results. The time duration of the syndrome diagnosis goes from the blood test authorization until the arrival of test results which can be days or weeks depending on the health systems (Irving et al. (2017); Minsalud (2015)). For patients at a high risk, the wait time is aggravated due to the occurrence of heart illnesses and diabetes mellitus (Jover et al. (2011)). In healthcare systems such as the one in Colombia, patients might not return to consultation for multiple reasons, including distance to hospitals, lack of enough providers, among other reasons, thus closing the possibility of starting the appropriate treatment.

The prediction models of metabolic syndrome without the need of taking a blood sample are essential for the community. Bypassing a blood exam is possible by using variables that can be obtained in a first-level medical care consultation at places such as a community health center. The medical staff can then follow up with a patient with some

risk factors of having the syndrome and, in consequence, a high probability of developing diabetes or heart disease at a lower cost because it is a computational model and can be implemented on a large scale easily. Moreover, no clinical exams will be performed to follow up, but only for patients that require a blood test to confirm a pathology.

Therefore, this chapter experiment the main prediction models of metabolic syndrome avoiding/bypassing a blood sample and using variables obtained in a first-level medical care consultation and published in the literature. As the main contributors to this research, it was tested the RAMAD methodology for the diagnosis of similar health conditions and a novel model. These models were compared with the prediction model proposed by the authors against data obtained from a study of metabolic syndrome performed on the Atlantic coast of Colombia in Barranquilla.

5.2 | Methodology

The absence of collaborative work is where the idea of the RAMAD methodology comes into play. Although it does not focus on collaborative work, it encourages a circular review and analysis process, with appropriate documentation. The methodology aims at building an excellent prediction model for the diagnosis of diseases using results obtained by the scientific community. This approach promotes continuous improvement through a circular and incremental methodology as was described in the chapter 2.

- **Review:** The chapter 4 was detailed the process of review with a window of seeking of 12 years counted retroactively from 2019 using four search engines: DBLP, IEEE XPLORE, ACM, and PubMed and that finally found 4 models of machine learning to diagnose the metabolic syndrome without doing a blood test.

- **Analyze:** The data used in this paper comes from a study of metabolic syndrome in Barranquilla performed in the second semester of 2012 by Universidad del Norte and was described in the chapter 2. In this chapter we use two criteria IDF and HMS to compare how that affects the prediction's performance of the model and that are detail in section of result.

- **Model:** The models found in the literature review chapter were used to perform various experiments and emulate their behavior using the theory of machine learning models and with the help of the data from the metabolic syndrome prevalence study described in chapter 2 emulate their performance.

- **Assess:** All the experiments were perform to assess each model using the hold out validation and/or random subsamplig validation to obtain the mean of each performance indicator detailed in the chapter 2.

- **Document:** This chapter gives specifications and detailed information about the data analysis performed, with a description such as mean and standard deviation by each feature. Such report must also include the parameters of the prediction models with their performance indicators and validation methods used for the experiments to allow for replication by other researchers.

5.3 | Results

The algorithms of ANN, decision tree, principal component logistic regression were performed and validated using MATLAB as well as the different types of validations such as hold out and random subsampling.

5.3.1 | Data Description

The analysis of the data from the patients who were diagnosed based on the IDF and HMS criteria is detailed in Table 5.2. The data is divided into two populations to explain the prevalence by gender, consisting of a total of 348(56.59%) women and 267(43.41%) men. Besides that, from the data, 252(40.98%) presented a diagnosis of MetS according to IDF, and 363(59.02%) were used as a control group.

On the other hand, with the HMS criteria as observed in Table 5.1, 262(42.6%) presented a diagnosis of MetS, and 348(57.4%) were used as control group. The analysis shows that only ten positive cases (7 men and 3 women) vary between the two criteria showing a possible association. It was checked with a Chi2 test resulting in $p \leq 0.0001$, evidencing a strong association between them. Nevertheless, both criteria were used for modeling.

Table 5.2: Relationship between IDF and HMS in the database.

Criteria	**IDF**		**HMS**	
Gender	No MetS(%)	MetS(%)	No MetS(%)	MetS(%)
Men	57.68%	42.32%	55.06%	44.94%
Women	60.06%	39.94%	59.2%	40.8%
Total	59.02%	40.98%	57.4%	42.6%

The age of the subjects is between 20 and 96 years, with an average of 43 years for women and 42 years for men. Also, the prevalence of MetS was 42.6% of the total sample divided into 44.94% among men and 40.8% among women for the criteria HMS and the prevalence of MetS was 40.98% of the total sample divided into 42.32% among men and 39.94% among women for the criteria IDF.

Table 5.3 shows the statistical description as average(m) and standard deviation (SD) of the variables: Age, Weight, Height, WHR, WSR, BMI, HC, WC, SBP and DBP of each group between healthy people, with MetS, No MetS, and the total using the IDF criteria. It also showed the Sex variable the number of women and men (women/men) by each group. The average value of all variables was higher in the MetS than in Non-MetS group (p value < 0.001).

Table 5.3: Statistic description of the total data using the HMS criteria.

Variables	**MetS**	**Non-MetS**	**Total**
AGE* m(SD)	47.62(17.49)	38.89(15.96)	42.61(17.17)
WC* m(SD)	99.81(11.33)	87.24(11.91)	92.59(13.21)
HC*: m(SD)	105.51(10.56)	93.73(12.50)	98.75(13.07)
BMI*: m(SD)	29.09(5.31)	25.26(4.74)	26.89(5.33)
WHR*: m(SD)	0.94(0.05)	0.93(0.09)	0.94(0.08)
WSR*:m(SD)	0.61 (0.67)	0.53 (0.74)	0.56 (0.79)
SBP*: m(SD)	128,52(18,46)	112,91(12,61)	119.55(17.19)
DBP*: m(SD)	78.48(11.13)	71.18(9.21)	74.29(10.69)
SEX (Women/Men)	(142/120)	(206/147)	(348/267)

Average(m); Standard deviation (SD); Women/Men; *(p<0.001)

5.3.2 | Experimenting with Models

This phase conducted experiments using the variables and techniques proposed in the analyzed works by Ivanovic(Ivanović et al. (2016)), Chen(Chen et al. (2014)), and other. The purpose was to build their performance indicators for the dataset of the population of the Atlantic coast of Colombia.

For the comparison, the experiments used only the data coming from the study of 615 subjects by Universidad del Norte. The variables used to compare were those found in all the articles, which are: Age, Sex, BMI, WC, HC, WSR, WHR, SBP and DBP, and other dichotomic variables. Healthcare professionals can obtain these variables at the first medical consultation.

On the other hand, healthcare professionals take into account the criteria shown in Table 5.1 for compliance with the risk factors, to diagnose the syndrome. The Waist Cir-

cumference (WC) and Blood Pressure (BP) variables are compared with their respective thresholds to become the positive (1) or negative (0) dichotomous value called WCD and BPD that represents the status normal and raised of the values of the waist circumference and arterial pressure, respectively. This approach transforms them into dichotomous variables, which details their values in Figure 5.1, where the dichotomous variables of systolic (SBPD) and diastolic blood pressure (DBPD) are also shown.

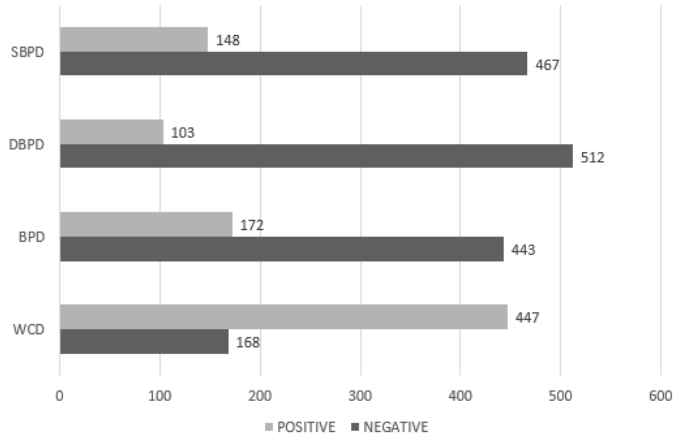

Figure 5.1: Dichotomous variables.

After that, the data was analyzed using the machine learning models proposed by each author. Kroon (Kroon et al. (2008)) proposed the regression tree technique, and the output is a dichotomous decision value for MetS. The authors used the BMI variables and two dichotomous variables WCD and BPD. These were used to replicate the experiment using algorithm shown in Appendix D.

However, the model proposed by Kroon (Kroon et al. (2008)) was built for patients 30 years old or younger. Therefore, the data was fixed to that range of age, resulting in a sensitivity 84.85%, specificity 53.85%, PPV 86.15%, NPV 51.22%, accuracy 77.78% and AROC 68.69% using IDF criteria. On the other hand, for practical purposing, we performed new experiments, but with all the ages, and the results improved the sensitivity of 80.27%, specificity 74,17%, PPV 82.92%, NPV 70,63%, accuracy 77,89% and AROC 76,78% using IDF criteria.

On the other hand, other authors, as (Murguía-Romero et al. (2013)) proposed an ANN with 25 hidden neurons, which has two neurons in the output layer, and the input has five neurons. One neuron by each one variable: WC, Sex, Height, Weight, BMI. This ANN was training using training data (70%) and testing data (30%) obtaining as resulted in a sensitivity 75.25%, specificity 58.62%, PPV 66.97%, NPV 68%, accuracy

67.39% and AROC 76.06% using IDF criteria.

From now on, it will be understood that all the ANN were used with the back-propagation configuration and will always have the same number of neurons as inputs variables in the input layer.

Chen (Chen et al. (2014)) used two data mining techniques to diagnose metabolic syndrome. First, it used Principal Component Logistic Regression (PCLR) with the variables SEX, AGE, BMI, WC, HC, WHR, SBP and DBP where it obtained the principal components of the training data and the testing data getting the following Equations (5.1)–(5.3).

$$PC_1 = -0.142Sex + 0.093Age + 0.230BMI + 0.253WC + 0.192HC + 0.192WHR + 0.178SBP + 0.160DBP \tag{5.1}$$

$$PC_2 = 0.045Sex + 0.402Age - 0.192BMI - 0.255WC - 0.184HC - 0.199WHR + 0.494SBP + 0.394DBP \tag{5.2}$$

$$PC_3 = 0.543Sex + 0.241Age + 0.278BMI + 0.001WC + 0.513HC - 0.429WHR + 0.046SBP - 0.207DBP \tag{5.3}$$

These were used to find the principal components through the 615 patients to test the model of logistic regression shown in the Equation (5.4) proposed by Chen (Chen et al. (2014)), using the training data.

$$Logit(P) = -1.809 + 1.722PC_1 + 0.276PC_2 + 0.403PC_3 \tag{5.4}$$

Then, $Logit(P)$ turns into the predictive variable Y_p with the Equation (**??**) mentioned in the Methdology section. So, we were getting performance indicators of sensitivity 59.02%, specificity 0%, PPV 100%, NPV 0%, accuracy 59.02% and AROC 49.86% using IDF criteria.

Chen (Chen et al. (2014)) then used the Equation (5.5) to normalize the variables and divided the dataset into two: training data (70%) and testing data (30%), to train and test an artificial neural network which has a back-propagation type configuration.

$$X_N = \frac{X - Xmin}{Xmax - Xmin} \tag{5.5}$$

This network has 5 hidden neurons with hyperbolic tangent sigmoid function and an output layer with a linear function. It is important to note that Chen (Chen et al. (2014)) did not publish the ANN configuration parameter, only the number of hidden

neurons and the activation function. Thus, it is necessary to build the ANN with the parameters shown in Table 5.4. This configuration was set by Kupusinac (Ivanović et al. (2016)) in its article to diagnose MetS.

Table 5.4: Parameter of the ANN (Ivanović et al. (2016)).

Parameter	**Value**
Training Function	Levenberg-Marquardt backpropagation
min_grad	10^{-10}
mu	10^{-3}
mu_dec	0.1
mu_inc	10
mu_max	10^{10}
HL function	hyperbolic tangent sigmoid
Out function	linear

Therefore, to replicate the Chen (Chen et al. (2014)) experiment of ANN with 5 hidden neurons, we decided to split the data into two parts: the training data (70% of the data) and the testing data (30% of the remaining data). With this step, the analysis shows a performance indicator of sensitivity 76,71%, specificity 71,67%, PPV 84,82%, NPV 59,72%, accuracy 75% and AROC 80,95% using IDF criteria.

Kupusinac (Ivanović et al. (2016)) proposed an optimization model to find the number of hidden neurons between 1 to 100 to obtain the maximum average of PPV and NPV, repeating the tests 100 times with the random subsampling validation and using the 6 variables SEX, AGE, BMI, WSR, SBP, and DBP. This approach creates two configurations of neural networks of 85 hidden neurons and 96 hidden layers.

When replicating the experiment, training and testing data are normalized to get the performance indicators with the same proportion of distribution in the data for comparing the performance indicators. After running the experiment, we obtained a mean sensitivity of 73.81%, specificity 65.89%, PPV 77.65%, NPV 60.18%, ACC 70.34%, and AROC 75.8%, for ANN with 85 hidden neurons. Moreover, ANN with 96 hidden neurons, we obtained a mean sensitivity 74.15%, specificity 66.03%, PPV 77.58%, NPV 60.8%, ACC 70.61% and AROC 76.64%.

To homogenize the experiments using ANN, Chen (Chen et al. (2004)) proposed a network of 5 hidden neurons with the random subsampling validation to get the performance indicators of sensitivity 76.47%, specificity 70.22%, PPV 80.41%, NPV 64.31% and AROC 81.75% using IDF criteria and In the same way, for the ANN of (Murguía-Romero et al. (2013)) with 25 hidden neurons, the results showed a sensitivity 72.64%,

specificity 63.78%, PPV 76.21%, NPV 58.47%, and AROC 77.13%.

5.3.3 | Data Analysis

A correlation analysis was also carried out between the variables that are related to obesity. Table 5.5 shows that Waist to Hip ratio (WHR) present a very low correlation with all the variables related to obesity. Also, the Waist to Stature Ratio (WSR), Hip Circumference (HC) and Waist Circumference (WC) present a high correlation with body mass index (BMI). This result reaffirms the importance of WC as a parameter to indicate the degree of obesity. It also shows why some criteria recommend it as a priority for the diagnosis of metabolic syndrome. Some criteria even consider it as eliminatory because if you do not meet this criterion at the same time, you could not have the syndrome according to the IDF criteria.

Table 5.5: Correlations between obesity related variables of Universidad del Norte data.

	BMI	**WC**	**HC**	**WSR**	**WHR**
BMI	1.00	0.78	0.79	0.79	0.06
WC	0.78	1	0.86	0.92	0.32
HC	0.79	0.86	1	0.84	−0.20
WSR	0.79	0.92	0.84	1	0.21
WHR	0.06	0.32	−0.20	0.21	1

After analyzing the correlations of the variables in the dataset and adding the dichotomic variables for the next analysis, we proceeded to select the variables that influence the best in diagnosing the metabolic syndrome to reduce the dimensions of the dataset to a subset.

For this purpose, we used Sequential Feature Selection searches for a subset of the features in the full model (Rückstieß et al. (2011)) with comparative predictive power being the variables described in Table 5.3 in conjunction with the dichotomous variables described in Figure 5.1.

Table 5.6: Correlation among HC, BPD and WCD.

	HC	**BPD**	**WCD**
HC	1.0000	0.3015	0.6459
BPD	0.3015	1.0000	0.2275
WCD	0.6459	0.2275	1.0000

A dataset consisting of the variables HC, WCD, and BPD is obtained and in an implicit way the variable gender because the threshold WC depends on gender. It also performed an analysis of correlation as shown in Table 5.6, which shows a very low correlation between them.

The selected variables from the dataset were divided into 70% for training and 30% for testing in an ANN. This technique was used due to a higher AROC when compared to the other techniques (decision tree and principal components logistic regression) to improve the diagnosis rate using only three variables that doctors can get in the first medical consultation and validating with random subsampling.

We chose three (3) hidden neurons due to methodology mentioned by (Boger and Guterman (1997); Karsoliya (2012); Panchal and Panchal (2014)) in which they established that the number hidden neurons should be 2/3 of input variables plus an output variable resulting in three.

We estimate the number of hidden neurons to contribute to research in the area of machine learning for the diagnosis of MetS without using biochemical variables and in a way, describe every detail of the process for experimentation by other researchers can continue investigating these models as well as Chen (Chen et al. (2014)).

This ANN configuration was trained and tested with normalized data to validate through random subsampling to obtain the performance indicators: sensitivity 76.17%, specificity 82.59%, PPV 90.54%, NPV 58.9%, ACC 77.41%, and AROC 87.36%.

As a summary, the performance indicators of each experiment is shown in the Tables 5.7–5.10, where the first two tables show results for the IDF criteria and the next two tables show results for the HMS criteria. Regardless of the diagnostic criteria, the behavior of data mining techniques shows that ANN of 3 hidden neurons is better compared to the previously proposed techniques. It seems that decreasing hidden neurons increases AROC.

Table 5.7: Performance indicator versus technique using hold out validation with IDF criteria.

	DT Kroon	**ANN25 Romero**	**PCLR Chen**	**ANN5 Chen**
SS	80.27%	75.25%	59.02%	76.71%
SP	74.17%	58.62%	0%	71.67%
PPV	82.92%	66.97%	100%	84.82%
NPV	70.63%	68%	0%	59.72%
ACC	77.89%	67.39%	59.02%	75%
AROC	76.78%	76.06%	49.86%	80.95%

Table 5.8: Performance indicator versus technique using random subsampling validation with IDF criteria.

	ANN96 Kupusinac	**ANN85 Kupusinac**	**ANN25 Romero**	**ANN5 Chen**	**ANN3**
SS	74.15%	73.81%	72.64%	76.47%	76.39%
SP	66.03%	65.89%	63.78%	70.22%	82.52%
PPV	77.58%	77.65%	76.21%	80.41%	90.24%
NPV	60.8%	60.18%	58.47%	64.31%	59.46%
ACC	70.61%	70.34%	68.79%	73.7%	77.46%
AROC	76.04%	75.8%	77.13%	81.75%	87.75%

Table 5.9: Performance indicator versus technique using hold out validation with HMS criteria.

	DT Kroon	**ANN25 Romero**	**PCLR Chen**	**ANN5 Chen**
SP	74.17%	59.46%	0%	68.52%
PPV	82.44%	71.96%	100%	84.40%
NPV	67.94%	57.14%	0%	49.33%
ACC	76.26%	65.76%	57.40%	70.11%
AROC	75.19%	75.48%	50.19%	78.97%

Table 5.10: Performance indicator versus technique using random subsampling validation with HMS criteria.

	ANN96 Kupusinac	**ANN85 Kupusinac**	**ANN25 Romero**	**ANN5 Chen**	**ANN3**
SS	71.62%	72.22%	69.44%	75.37%	75.08%
SP	66.95%	66.25%	63.73%	72.54%	80.15%
PPV	77.4%	76.05%	75.74%	81.38%	87.17%
NPV	58.67%	60.80%	55.34%	64.21%	59.94%
ACC	69.33%	69.42%	68.88%	73.91%	75.35%
AROC	74.94%	74.79%	74.8%	81.74%	85.13%

However, it is important to note that each ANN has a different dataset of variables, and so, these are different each only ANN85, and ANN96 has the same variables.

Tables 5.7–5.10 shows that two techniques or configurations were better in some performance indicators. ANN using 8 variables (AGE, SEX, WHR, BMI, HC, WC, SBP, DBP) with 5 hidden neurons, which has a sensitivity of 80.27%, NPV 70.63%, and ACC 77.89% higher when compared with other machine learning models. However, ANN with 3 hidden neurons using 3 variables (WCD, DAP, HC) and SEX, which was im-

plied, is the best due to the performance indicator of specificity 82.59%, PPV 90.54% and specially AROC 87.36%.

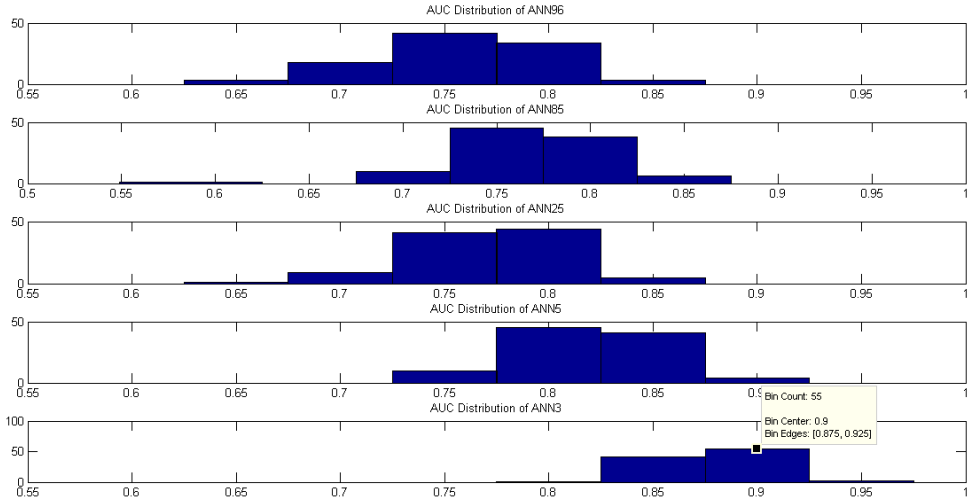

Figure 5.2: AROC distribution for each ANN to diagnose using the IDF criteria.

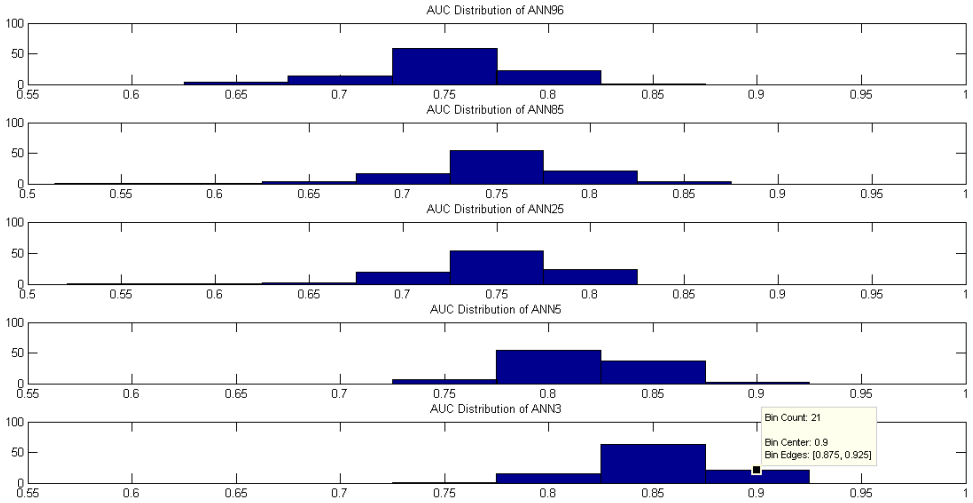

Figure 5.3: AROC distribution for each ANN to diagnose using the HMS criteria.

The previous results allow classifying patients with MetS (i.e., the sensitivity) very selectively. This selective prediction is possible because the healthcare professionals always presume that every patient is healthy, and this algorithm can guarantee that

when a patient with MetS is identified, it is because they have a high probability (approximately 90.54%) that they suffer from it. Concerning specificity, the ANN obtained 82.59%, being superior to the other models, indicating that the ANN is an excellent tool to rule out the probability of the positive diagnosis of the syndrome, because of every 100 healthy patients, the ANN confirms 83 approximately healthy people.

On the other hand, as regards to AROC, the ANN technique obtained 87.36%, being superior when compared to others. This result means that an individual randomly selected from the group of patients has a test value higher than one randomly selected from the healthy group at 87.36% of the time considered excellent according to the dissertation of Hosmer and Lemeshow (D. W. Hosmer and Lemeshow (2004)).

The mean was used to quantify the value of the highest probability as IDF as HMS criteria diagnostic because we assume that random subsampling would generate a normal probability distribution. We plotted the distribution of the AROC probabilities, and this was considered as a discriminant value to choose the best classifier among ANN96, ANN85, ANN25, ANN5, and ANN3. Figures 5.2 and 5.3 show that the average does not reflect the highest AROC probability rate. The ANN3 has, in its zone of occurrence, higher AROC levels unified in the range between 0.875 to 0.925 with a probability of 55%, and median 87.43% to IDF. The same happens with the HMS criteria, which presents higher AROC levels unified in the range 0.875 to 0.925 with a probability of 22%, and median 85.36%. Hence, HC, BPD, and WCD input variables in ANN with 3 hidden neurons have an excellent performance concerning the other proposed models with the dataset collected by Universidad del Norte in Barranquilla.

5.4 | Discussion

Metabolic syndrome is a constellation of several risk factors of developing cardiovascular disease and type 2 diabetes that requires a blood test for diagnosis. However, with the technique of data mining, the scientific community has built models to obtain data that can get through first-level medical attention without the need of a blood test. This paper has shown a possible association between IDF and HMS criteria that is checked with a Chi2 test resulting in $p \leq 0.0001$, evidencing a strong association between them which was explained by the difference of only 10 positive cases (3 women and 7 men) of MetS. Also, according to the review phase of the proposed methodology, the most used criteria in the models used to diagnose MetS without blood test were the IDF and in second place was the ATP III, which have some small differences in the levels of some variables. However, since 2009, the specialists have had the consensus of

Harmonized Diagnosis (HMS) to unify the diagnosis of the metabolic syndrome, which a previous paper already used (Murguía-Romero et al. (2013)). Thus, the models explained used IDF and HMS criteria to compare the results with other future models.

The analyzed learning models allow the diagnosis of the metabolic syndrome with an AROC higher than 70% without using a blood sample, early avoiding or delaying its appearance because some risk factors such as obesity can be modified. It is interesting to note that the ANN technique has better precision than PCLR and DT. However, the latter provides rules and equations that researchers can implement at a lower computational cost. They also generate a priori knowledge about the relationships of the non-invasive variables with the metabolic syndrome due to their white-box model.

Several models used variables relative to the obesity such as Body Mass Index, Waist Circumference, Hip Circumference, Waist to Hip ratio, and Waist to Stature ratio. During the correlation analysis step of the data analysis phase, we found a good correlation between them except with Waist to Hip ratio. These variables, when processed together, can generate a problem that commonly exists in the medical statistical literature, called multicollinearity. A way to avoid the multicollinearity is to use variable transformation. Waist Circumference WC becomes a Dichotomous variable (WCD) to decrease the correlation with Hip Circumference.

The data indicated that the ANN model of three hidden neurons using WCD, BPD, and HC according to its AROC of 87.75% to IDF and 85.12% to HMS are excellent to discriminate the data. The results are excellent in comparison to principal component logistic regression and decision trees and the other ANN model. All were compared based on the biomedical dataset of 615 people of the Atlantico, Colombia region, of whom 42.6% had metabolic syndrome.

Therefore, it is recommended to make more comparisons with a larger dataset to strengthen the thesis that the ANN model is superior for the diagnostic of metabolic syndrome without using a blood test and so, we only take measurements of triglycerides, HDLC and glycemia in necessary cases. Thus, this project only used variables that can be obtained in a medical consultation at the first level of health care or community health center.

The two decision tree and ANN techniques have in common that they diagnose through three variables of which two are common and are WCD and BPD and demonstrate the importance of dichotomous variables when diagnosing diseases and support why they are recognized as risk factors.

Researchers could replicate the methodology and its application presented in this paper in other areas where data collection from patients is happening. Such projects include those where researchers obtain biomedical data from trials or the immersion

of health monitoring sensors in the world (for example, smart city, smart wearables, among others). These applications, together with the technique of data mining, can find patterns to make decisions and find many solutions. Such is the case of the diagnosis and monitoring of pathology development, and in consequence, the improving of the well-being of people and disease prevention with the help of these learning models.

This chapter proposed a methodology to find a model to diagnose diseases under certain conditions, in this case, is MetS without a blood test using variables obtained in first-level medical attention were the result of the effective ANN model with three hidden neurons and only three variables. Therefore, we recommend the methodology to get a model generalized or outstanding model. Furthermore, it is very important and necessary to test with more data from different populations.

RAMAD methodology offers the advantages to a range of entities, such as doctors, patients, health systems and communities. Doctors have the advantage of obtaining the method with better performance to be able to predict diseases that depend on variables obtained with non-invasive procedures, such as the metabolic syndrome. On the other hand, the patients with an early diagnosis can go on with the treatment immediately managing to reduce their risk factors that are related to diabetes and heart disease.

As for the healthcare system, it will be able to do the proper monitoring of diseases in patients who live in remote areas by predicting their status, before the disease becomes intractable and deadly. So, it manages to reduce the cost of health care monitoring and management caused by these diseases and the delaying of the treatment. It can also help generate new knowledge about patients with MetS diagnosed in a non-invasive way by discovering new treatments that, when implemented, will reduce the risk of developing diseases resulting from MetS.

By having RAMAD methodology, the community will be able to acquire new non-invasive diagnostic methods that will allow an increase in the level of health and well-being due to the fact that people will measure the progress of their health status with the help of technological applications that will keep them motivated.

In the very near future, the RAMAD methodology can be improved using more precise methods such as deep learning with a greater number of data captured in real time with the implementation of smartcity and big data where the end user, that is, the patient can use the emerging technologies in the field of health care, clinical score, m-health, e-health (Vashist et al. (2014)). Additionally, the next-generation point of care (POC) technologies will benefit from advances in data mining methodology for the improvement of the parameters of which some are used and the increase in the number of users that require the use of this technology (Vashist and Luong (2019),Vashist et al. (2015)).

In general, this is a sufficiently detailed study and that reports a novel methodology which can be applied not only to MetS, but also to a variety of other diseases, such as cardiovascular diseases, diabetes, etc. Moreover, this opens the possibility to enable the development of generalized models for such diseases.

5.5 | Summary

Metabolic syndrome is diagnosed according to the most up-to-date methods, and all healthcare professionals agree on using five criteria, three of which are obtained through blood analysis. Healthcare providers then collect the results and apply the criteria through the decision threshold according to each method. Although using principal component logistic regression, decision trees for the processing of data obtained from non-biochemical variables, it has been proven that metabolic syndrome can be diagnosed with an acceptable discrimination ability because of an AROC ≤ 0.8. Being the learning model of ANN with 3 hidden neurons obtained a higher AROC of 87.75% to IDF and an AROC of 85.12% to HMS that are excellent. Therefore, ANN with three hidden neurons using WCD, BPD, and HC is an excellent model for identifying metabolic syndrome and helps to reduce the time of treatment initiation, which allows for the development of simple prevention strategies for the subject at risk for type 2 diabetes and cardiovascular disease.

The results of this chapter reflects the importance of monitoring blood pressure, hip circumference, and waist circumference for diagnosing the metabolic syndrome. The results also reflect the importance of dichotomous variables in medical decision making. Caregivers can obtain these variables through tools usually essential in the first-level medical attention such as a tape measure and a sphygmomanometer.

These results come from the application of the proposed methodology, which highlights the importance of continuously reviewing new findings, applying changes to the model, and adjusting accordingly. Therefore, the contribution of this chapter is not only the model but a continuous methodology to get a generalized model using data mining techniques which should attract other researchers to contribute and continue with this work to achieve a generalized model of MetS. Researchers could apply such methodology to other diseases to construct generalized models for those diseases.

6

Framework for Prediction of Metabolic Syndrome Types without doing a Blood Test

This chapter explain the design and implementation of a framework to predict the metabolic syndrome a without doing a blood test based on machine learning. Moreover, it proposes a segmentation of metabolic syndrome to predict the types of metabolic syndrome according to HMS criteria building models for each type base on machine learning in order to compare the general metabolic syndrome with each type.

6.1 | Introduction

Previously, the different criteria for diagnosing MetS have been described. But, independently of the criteria to diagnose MetS, the diagnosis uses five factors for example, the HMS criteria that is the most updated as shown Table 6.1. Doctors get two of those factors (Waist Circumference (WC) and Blood Pressure (BP) in medical consultations and a community setting. Invasive tests are required to know the value of triglycerides, HDL-C, and fasting plasma glucose present in the patient's blood. So, the time of treatment initiation can vary according to the health system. The delay between initial consultation, a blood test to measure the triglycerides, fasting blood sugar, and HDL-C levels, plus a diagnostic consultation, can add several days or weeks (Irving et al. (2017); Minsalud (2015)) creating a problem to diagnose early.

Table 6.1: Definition of the HMS criteria of the MetS according to HMS (Data from (Aschner (2010))).

Risk Factors	**HMS Criteria**
Central Obesity	Waist Circumference (WC) population and country specific
Triglycerides (TG)	$\geq$150 mg/dL
Fasting Plasma Glucose (FPG)	$\geq$100 mg/dL
High-Density Lipoprotein Cholesterol (HDL-C)	<40 mg/dL in males <50 mg/dL in females
Blood Pressure	Systolic $\geq$ 130 mmHg and/or Diastolic $\geq$ 85 mmHg
Diagnostic criteria	Three risk factors

This time delay in obtaining the results, particularly for patients in remote locations, might be sufficient time in some cases to worsen or aggravate the patient's health conditions due to the occurrence of a stroke diabetes(Jover et al. (2011)). It is useful to predict early MetS to avoid or delay the onset of some illnesses already mentioned.

Many researchers have proposed solving the problem without doing a blood test using machine learning techniques, such as Kroon et al. (Kroon et al. (2008)), Hsiung et al. (Hsiung et al. (2015)) and others. However, when diagnosing MetS, doctors always check the triglycerides, fasting plasma glucose, and HDL-C values to recommend a specific treatment to prevent diabetes or coronary diseases. They also want to know the possible cause since it allows a better decision to plan patient treatment (Alberti et al. (2009); Alshehri (2010)).

Therefore, this chapter proposes segmenting MetS into various types to identify the risk factors that produce it and use machine learning to predict them early without making a blood test and comparing each MetS type with the traditional MetS. Moreover, this chapter analyzes four approaches for improving the AROC for the different MetS types. The first approach uses the ANN technique; the second approach uses an ensemble prediction algorithm as the Random undersampling Boosted tree (RusBoost) ensemble. The third approach uses an oversampling technique to create more data and then applying ANN. The last approach uses the dataset with oversampling and RusBoost.

The objectives of this chapter are the following:

- Achieve a mathematical representation to diagnose MetS using HMS criteria.
- Propose a segmentation of MetS using HMS criteria.
- Develop a framework to predict the different MetS types according to HMS criteria using a set of variables that doctors can obtain using non-invasive methods in a first consultation.
- Evaluate two machine learning techniques using performance indicators for each MetS type.

6.2 | Methodology

To continue improving the framework, this chapter uses the methodology called RAMAD(Barrios et al. (2019)) and is composed of the stages review, analysis, model, assess, and document with the in order to develop the project and was explained in detailed in chapter 3. In chapter 5 was found that artificial neural network is better in comparison of decision tree and logistic regression to diagnose of MetS without doing a blood test and so, the artificial neural networks will be used to build a model for the framework.

The execution of the steps can be updated. Researchers could do so by repeating the methodology from the first phase. In this way, they could add new models from the literature. This circular approach ensures that researchers keep improving their models. If they follow the documentation phase correctly, new researchers could improve the original model's prediction capabilities.

6.2.1 | Review

The Review phase performed a specialized search in the following databases: DBLP, IEEE, and ACM for the relationship with Computer Science, and Pubmed for its relationship with healthcare research. I had used a window of 12 years since 2008. The keywords used were "SEGMENTATION" OR "TYPES" OR "PREDICTION" OR "ANN" AND ["METABOLIC SYNDROME" AND "WITHOUT BLOOD TEST"] for Query1: IEEE, Query2: DBLP, Query3: ACM, and Query4: PUBMED to know all the projects related to keywords in these search engines dedicated to the field of engineering. The results of the queries (QUERY1: 92, QUERY2: 51, QUERY3: 82, and QUERY4: 131) provided an extensive list of journal articles and conferences. However, not all the documents are directly related to the segmentation of MetS since DBLP, IEEE XPLORE, ACM, and PUBMED

delivered articles on machine learning and computer science at a general level, for example, the topic image segmentation.

Then, the delivered list was filtered, evaluating its relationship with the segmentation of MetS or types of MetS. It excluded repeated articles (mirror articles). Afterward, it performed a manual inspection and then filtered by the criteria established on the following questions.

- Could authors predict the metabolic syndrome types or segmentation without a blood test? Y/N.
- What metabolic syndrome diagnostic criteria did the authors use? e.g., ATP II, IDF, HMS, or other criteria recognized.
- What ANN configuration did the authors use?
- What validation method did the authors use? e.g., hold out, random subsampling, and others.
- What performance indicators did the authors use? e.g., Sensitivity, Area Under the ROC curve, Specificity.

The manual inspection did not find anything about the segmentation of MetS using only variables obtained in a medical consultation such as anthropometric and clinical or history variables. However, in the chapter of the review literature found three articles that diagnose MetS using ANN and anthropometric and clinical variables such as Age, Sex, Weight, Height, Waist Circumference (WC), Hip Circumference (HC), Waist to Hip ratio (WHR), Waist to Stature (WSR), Body Mass Index (BMI), Body Fat Percentage (BFP), Systole Blood Pressure (SBP) and Diastole Blood Pressure (DBP). These variables will be used to build the model to predict the MetS types without doing a blood test, that is, without using biochemical variables.

As a summary, Table 6.2 shows an overview of the variables that are used explicitly in the study and other variables that were not used (implicitly) but are necessary to construct the explicit variables. It also shows the prediction models used for the MetS prediction without doing a blood sample of each article in the literature.

Diagnosing MetS using non-biochemical variables is an approach that implies not taking blood samples. This approach can help doctors make early decisions about MetS. However, doctors always need to know what risk factors are present in the patient diagnosed with MetS to start treatment early to decrease the probability of heart disease or diabetes mellitus type 2. Moreover, to date, no study has evaluated the MetS types or

Table 6.2: Variables and hidden neurons of ANN used by the authors found in the review.

Authors	**Murguia-Romero**	**Chen**	**Kupusinac**
Age		E	E
Sex	E	E	E
WG:Weight	E	I	I
HG:Height	E	I	I
WC: Waist circumference	E	E	I
HC: Hip circumference		E	
WHR: Waist to Hip ratio		E	
WSR: Waist to Stature Ratio			E
BMI: Body Mass Index	E	E	E
SBP: Systolic blood pressure		E	E
DBP: Diastolic blood pressure		E	E
Hidden neurons	25	5	85 and 96

E: Explicit use of variable; I: Implicit use of variables

the segmentation of MetS from a perspective of machine learning. This situation may be due to the lack of a model of segmentation of MetS that explains the different MetS types.

6.2.2 | Analysis

Continuing with the study of the metabolic syndrome that was detailed in the chapters 2 and 5. But, now including other variables that in many articles have supported the association between MetS and the percentage of body fat obtained with the bioimpedance technique(Navarro et al. (2016); Rodríguez et al. (2017)). However, according to the objectives stated above, the variables should be obtained from the medical consultation

data where, in general, the first level assistance office does not have a body fat measurement device.

Therefore, the measurement of Body Fat Percentage (BFP) was performed with the following equations depending on gender for men Eq. 6.1 and for women Eq. 6.2. According to an analysis of the authors Lean, Han, and Deurenberg(Lean et al. (1996)), equations 6.1 and 6.2 have the largest prediction power to measure BFP.

$$BFP(\%) = 0.567WC(cm) + 0.101Age(year) - 31.8 \tag{6.1}$$

$$BFP(\%) = 0.439WC(cm) + 0.221Age(year) - 9.4 \tag{6.2}$$

The research detailed in the chapters 2 and 5 also obtained information about the respondents' health history that was recommended by several authors(Chen et al. (2014); Hsiung et al. (2015); Kroon et al. (2008)). According to the variables of the study, we analyze the history of obesity with the variable obtained from the following question: "Have any health professionals diagnosed you with overweight or obesity?." Recent studies have shown a strong association of the previous obesity diagnosis to the risk of heart failure (HF)(Fliotsos et al. (2018)). Therefore, we tabulate the discrete values of the Previous Obesity Diagnosis (POD) variable in the Results section.

6.2.3 | Model

In this section proposes a model of segmentation of MetS obtaining several types of MetS for HMS criteria. Then, it shows an abstract overview of a framework for implementing the proposed method to predict each type of MetS without doing a blood test. First, it analyzes the Mathematical representation to diagnose MetS.

6.2.3.1 | Mathematical representation to diagnose MetS

The MetS, according to different organizations traditional (HMS, IDF, ATP III, and among others), must be diagnosed when it meets at least three risk factors, that is, 3, 4 or 5 dichotomous variables of the risk factors $W, P, G, H, and\ T$.

- W: Represents the status normal(0) or raised(1) of the dichotomous values of the waist circumference
- P: Represents the status normal(0) or raised(1) of the dichotomous variable of the blood pressure

- G: Represents the status normal(0) or raised(1) of the dichotomous variable of the fasting plasma glucose
- H: Represents the status normal(0) or lowed(1) of the dichotomous variable of the HDL-C
- T: Represents the status normal(0) or raised(1) of the dichotomous variable of the triglycerides

It is essential to keep in mind that it was obtained the dichotomous variable of blood pressure (P) by making a logical OR operation between the dichotomous variables of Systolic Blood Pressure (SBPD) and Diastolic Blood Pressure (DBPD), as shown in the Eq. 6.3.

$$P = SBPD \mid DBPD \tag{6.3}$$

If we base the verification of a risk factor through the criteria shown in Table 6.1 and then it could represent each risk factor as a dichotomous variable, where one is positive, and 0 is not positive or negative. Moreover, it considered F =W, P, G, H, T as a set of risk factors. So, the diagnosis with the criterion HMS can be represented mathematically through the sum of combinations without repetition of 3, 4, and 5 from 5 dichotomous variables of the risk factors, as shown in Eq. 6.4 with its result of 16 combinations of the risk factors of MetS.

$$\sum_{p=3}^{5} \binom{5}{p} = \binom{5}{3} + \binom{5}{4} + \binom{5}{5} = 16 \tag{6.4}$$

These are the combinations to diagnose the metabolic syndrome as positive with the HMS criteria according to Truth Table 6.3. It builts from the Boolean perspective represented by the logical sum (logic OR) of products(Floyd (2010)) labeled numerically through the format $(WPGHT)$ base 2 converted to base 10 as shown Eq. 6.5 .

$$MetS_{HMS} = \sum_{5}(07, 11, 13, 14, 15, 19, 21, 22, 23, 25, 26, 27, 28, 29, 30, 31) \tag{6.5}$$

Next, it optimizes the Eq. 6.5 and minimize it using the Quine-McCluskey algorithm. The detailed solution is in Appendix E and we checked the technique with Karnaugh Map as well, obtaining the Eq. 6.6 in the format $WPGHT$.

$$MetS_{HMS} = WPT + WPH + WPG + WGT + WGH + WTH + PGT + PGH + PHT + GHT \tag{6.6}$$

Table 6.3: Truth table of all the combinations of the risk factors of the MetS according to the HMS criteria.

n	W	P	G	H	T	MetS
0	0	0	0	0	0	0
1	0	0	0	0	1	0
2	0	0	0	1	0	0
3	0	0	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	1	0
6	0	0	1	1	0	0
7	0	0	1	1	1	1
8	0	1	0	0	0	0
9	0	1	0	0	1	0
10	0	1	0	1	0	0
11	0	1	0	1	1	1
12	0	1	1	0	0	0
13	0	1	1	0	1	1
14	0	1	1	1	0	1
15	0	1	1	1	1	1
16	1	0	0	0	0	0
17	1	0	0	0	1	0
18	1	0	0	1	0	0
19	1	0	0	1	1	1
20	1	0	1	0	0	0
21	1	0	1	0	1	1
22	1	0	1	1	0	1
23	1	0	1	1	1	1
24	1	1	0	0	0	0
25	1	1	0	0	1	1
26	1	1	0	1	0	1
27	1	1	0	1	1	1
28	1	1	1	0	0	1
29	1	1	1	0	1	1
30	1	1	1	1	0	1
31	1	1	1	1	1	1

As observed in Eq. 6.6, the tripartite variables that we call MetS types are always necessary for a diagnosis of the traditional MetS. It requires at least one of these WPT, WPH, WPG, WGT, WGH, WTH, PGT, PGH, PHT, and GHT to be positive according to the HMS criteria, as detailed in Table 6.4.

The term traditional MetS or general MetS will be used to separate it from the MetS

types proposed in this chapter. For practical purposes, it developed a framework to predict the MetS types based on the Eq. 6.6 representing the HMS criteria using machine learning techniques.

6.2.3.2 | Framework to predict the MetS types

Figure 6.1 shows an abstract overview of the implementation of the proposed method by using a framework divided into three stages.

Table 6.4: Types of MetS according to the HMS criterion

Type	Diagnostic of MetS
WPT	Increased Waist Circumference , Blood Pressure, and Triglycerides levels
WPH	Increased Waist Circumference , Blood Pressure, and reduction of HDL-C levels
WPG	Increased Waist Circumference , Blood Pressure, and Fasting Plasma Glucose levels
WGT	Increased Waist Circumference , Fasting Plasma Glucose, and Triglycerides levels
WGH	Increased Waist Circumference , Fasting Plasma Glucose, and decreased HDL-C levels
WTH	Increased Waist Circumference , Triglycerides, and decreased HDL-C levels
PGT	Increased Blood Pressure, Fasting Plasma Glucose, Triglycerides levels
PGH	Increased Blood Pressure, Fasting Plasma Glucose, and decreased HDL-C levels
PHT	Increased Blood Pressure, Triglycerides and decreased HDL-C levels
GHT	Increased Fasting Plasma Glucose, Triglycerides and decreased HDL-C levels

1. Extraction, Transformation, and Load (ETL)

   In this stage, we collected the data from a population of 615 subjects who authorized taking a blood sample to measure the values of triglycerides, HDL-C, and fasting plasma glucose. Moreover, the study recorded the anthropometric and clinical variables such as Age, Sex, Weight, Height, Waist Circumference (WC), Hip Circumference (HC), Systole Blood Pressure (SBP), and Diastole Blood Pressure (DBP).
   Later, through the transformation process, we obtained Body Mass Index, Body

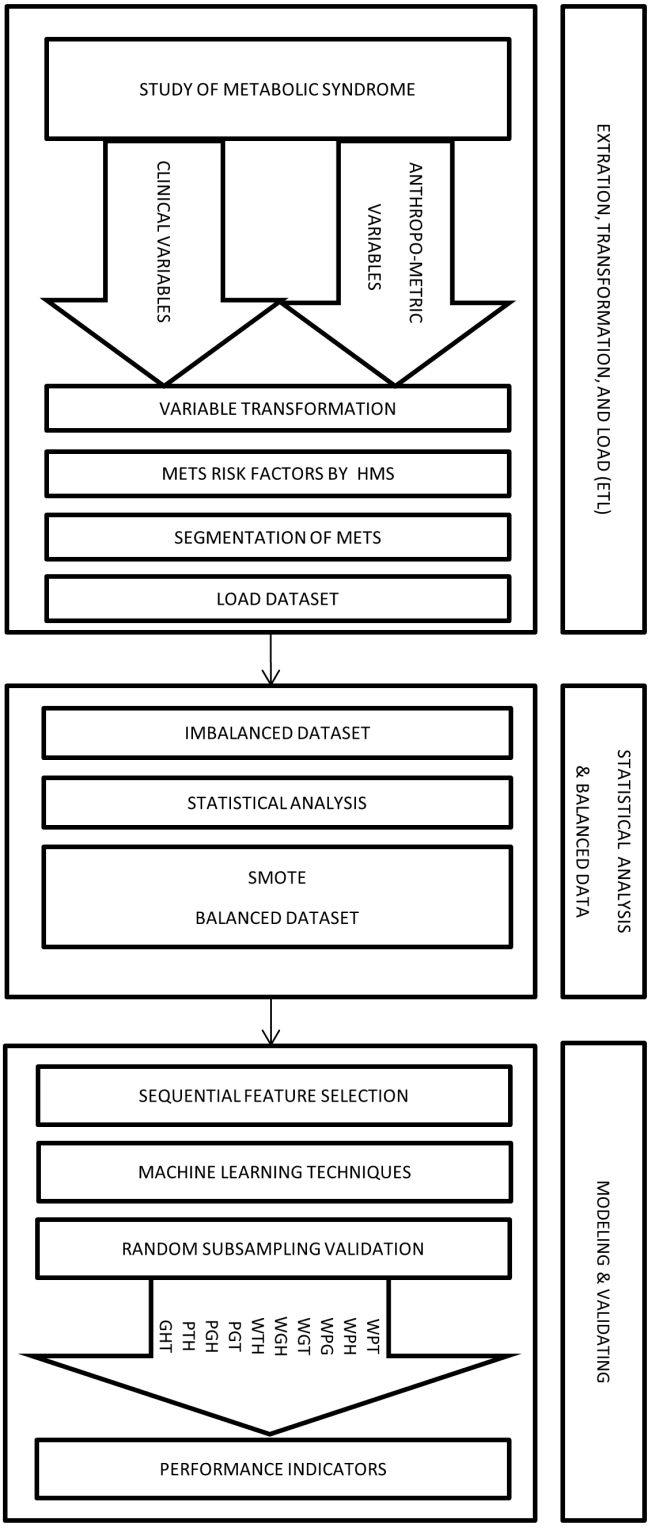

Figure 6.1: Framework to predict the types of MetS by HMS criterion using non-biochemical variables

Fat Percentage, Waist Hip circumference ratio, Dichotomous Blood Pressure Systolic, Dichotomous Diastolic Blood Pressure, Dichotomous Blood Pressure, Dichotomous triglycerides, Dichotomous fasting blood sugar, Dichotomous HDL-C, and Dichotomous Waist circumference among others.
Afterward, we used dichotomous values of the HMS criteria' risk factors to build the different MetS types obtained from the segmentation process explained in the previous subsection. We obtained the output variables WPG, WPH, WPT, WGH, WGT, WTH, PGT, PGH, PHT, and GHT. So, all anthropometric and clinical data was loaded in a dataset of 615 records.

2. Statistical analysis and dataset balancing

   In this stage, we began with a dataset containing 615 people with samples of biochemical variables with their respective diagnostic of MetS. Then, we did a descriptive statistical analysis of the dataset, finding that some types of MetS were imbalanced, as shown in the Results section.
   This problem was caused by the low prevalence of the risk factor for fasting blood glucose in the study population. This low prevalence is expected in a study of MetS(Perveen et al. (2019)). It resolved this imbalance by using a data balancing technique, such as the Synthetic Minority Over-sampling Technique (SMOTE) implemented by WEKA. We created synthetic data to get a balanced dataset of 799 records (615 plus 184 synthetic data) and a better distribution of risk factors of MetS, thus improving the quality of discrimination.

3. Modeling

   In this stage used an algorithm to select the necessary non-biochemical features as well as used Sequential Feature Selection in Matlab to achieve the maximum discrimination in both datasets (imbalanced and balanced) of the proposed model's output variables. For the following step, it used several Multilayer Perceptron (MLP) ANN to predict the MetS general, and each MetS type: WPG, WPH, WPT, WGH, WGT, WTH, PGT, PGH, PTH, and GHT.

   Each ANN has several hidden neurons that have functions, such as the hyperbolic tangent sigmoid function and an output layer with a neuron. The neuron has a function that can be a log-sigmoid function (Bishop (2006); Witten and Frank (2005)).

   It should be noted that there are no hard and fast rules for the number of hidden neurons. These hidden neurons can be calculated or found empirically and are

highly dependent on the problem and the dataset (Andrea and Kalayeh (1991)). However, we used the methodology mentioned by (Boger and Guterman (1997); Karsoliya (2012); Panchal and Panchal (2014)) and described in Equation (6.7), where the number of hidden neurons (NHN) can be 2/3 of the input variables plus an output variable.

$$NHN = \frac{2(Input\ variables)}{3} + Output\ variables \tag{6.7}$$

It used Equation (6.7), to estimate the number of hidden neurons to contribute to research in the area of machine learning for the prediction of MetS without using biochemical variables and in a way, describe every detail of the process for experimentation by other researchers can continue investigating these models as well as Chen (Chen et al. (2014)) that used other equation to calculate the hidden neurons.

Another machine learning technique used to predict the MetS types was the ensemble Random undersampling Boosted tree (RusBoost)(Mounce et al. (2017)) with the configuration showed in Table 6.5 due to the data from the MetS study is usually imbalanced.

Table 6.5: RusBoost Configuration

Learned Type	Decision tree
Maximum number of splits	20
Number of learners	30
Learning rate	0,1

In summary, we used two machine learning techniques ANN and RusBoost to design the models and validating with the performance indicators using random subsampling described in the next subsection.

6.2.4 | Performance Indicators and Model Assessment

This chapter used a dataset summarized by means, standard deviations, and percentages and found the prevalence of MetS and analyzed each variable of the dataset in two groups (MetS and Non-MetS) assessed with t-tests and Chi2 tests using the SPSS statistical software, version 23 for Windows.

For the validation of the model, it used random subsampling to compare the classification models to predict the MetS with each MetS type (WPG, WPH, WPT, WGH,

WGT, WTH, PGT, PGH, PTH, and GHT) without doing a blood test. For this purpose, it used indicators to evaluate their capacity for discrimination, such as Sensitivity (SS), Specificity (SP), False Negative Rate (FNR), False Positive Rate (FPR), AROC (Perveen et al. (2019); Witten and Frank (2005)).

The classifiers used to predict diseases prioritize the rate of type 1 and 2 errors known as False Negative Rate (FNR) and False Positive Rate (FPR), which are the complement of SS and SP, respectively. The type 2 error (false negative) is harmful because a patient with metabolic syndrome can be diagnosed as a patient without the syndrome, affecting their health since they did not start adequate treatment.

6.2.5 | Document

In the following results section, it presents the analysis and document the whole process to be reused later by other authors to improve the proposed model. It records all the models' parameters and verify that it is detailed so that the project and experiments can be replicated.

6.3 | Results

This section describes the results obtained from the experiments to find a data description of the variables and the performance indicators described previously to predict the traditional MetS and each one of the MetS types based on the dataset from a MetS study conducted by the Universidad del Norte.

6.3.1 | Data description

The Universidad del Norte conducted the MetS study with a sample of 615 patients split into 348 women and 267 men between 20 and 96 years old. The study used blood tests to obtain the biochemical variables of triglycerides, fasting plasma glucose, HDL-C. The results are separated into three groups: MetS, Non-MetS, and total, as shown in Table 6.6.

The biochemical variables of the MetS group and the Non-MetS group have different statistical significance. Together with the waist circumference and systolic and diastolic blood pressure, these variables are necessary to predict the MetS. The study found that the total prevalence rate was 42.60% divided into 44.94% for men and 40.8% for women.

Moreover, healthcare professionals collected anthropometric and clinical variables such as Age, Sex, Weight, Height, WHR, HC, WC, SBP, and DBP in each patient. Other

Table 6.6: Statistic description of the biochemical variables.

Variables *	MetS m(SD)	No MetS m(SD)	Total m(SD)	p
TG	216.94 (112.8)	121.84 (63.05)	160.81 (98.67)	<0.001
GL	97.33 (38.82)	84 (19.56)	89.47 (29.74)	<0.001
HDL-C[W]	38 (8.22)	46.97 (13.54)	43.39 (12.49)	<0.001
HDL-C[M]	36.11 (11.1)	43.32 (11.63)	40.27 (11.93)	<0.001

TG: Triglyceride; GL: Fasting Blood Glucose; M: Men; W: Women; Average(m); Standard deviation (SD); * (mmol/L).

Table 6.7: Statistical description of the study variables for the total data.

Variables	MetS m(SD)	No MetS m(SD)	Total m(SD)	p
Age (year)	47.62 (17.49)	38.89 (15.96)	42.61 (17.17)	<0.001
WC (cm)	99.81 (11.33)	87.24 (11.91)	92.59 (13.21)	<0.001
HC (cm)	105.51 (10.56)	93.73 (12.50)	98.75 (13.07)	<0.001
Weight (Kg)	79.08 (17.11)	66.59 (13.81)	71.71 (16.43)	<0.001
Height (m)	1.64 (0.09)	1.62 (0.09)	1.63 (0.09)	0.068
BMI (Kg/m)	29.09 (5.31)	25.26 (4.74)	26.89 (5.33)	<0.001
WHR*	0.94 (0.05)	0.93 (0.09)	0.94 (0.08)	<0.001
WSR*	0.61 (0.67)	0.53 (0.74)	0.56 (0.79)	<0.001
BFP (%)	38.64 (8.46)	30.86 (10.23)	34.05 (10.28)	<0.001
SBP (mmHg)	128,52 (18,46)	112,91 (12,61)	119.55 (17.19)	<0.001
DBP (mmHg)	78.48 (11.13)	71.18 (9.21)	74.29 (10.69)	<0.001

*(cm/cm);Average(m); Standard deviation (SD).

variables were calculated, such as BFP and BMI. Table 6.7 shows the statistical description of the total data of the study with the variables between people with MetS, No MetS, and total. These variables were measured, given that several studies (Chen et al. (2014); Ivanović et al. (2016); Murguía-Romero et al. (2013)) suggested to take into account these variables to predict the MetS without a blood sample. Doctors can obtain those variables at the first medical consultation. We can observe that the average of each of these variables obtained from patients with MetS is higher than that of patients without MetS and present a statistical significance in the groups MetS and Non-MetS that evidence a difference between groups. On the contrary, the Height variable's behavior in the two groups demonstrates a ($p = 0.068$) very low probability of difference.

Other variables that should be taken into account to predict the MetS are those found in the clinical history such as the Previous Obesity Diagnosis (POD) due to the relationship with the occurrence of coronary heart disease (Fliotsos et al. (2018); Perveen et al. (2019)). Therefore, in this MetS study, the researchers asked patients about their his-

tory of a previous obesity diagnosis and found that 42.37% were MetS and 23.23% were Non-MetS with a significant difference of $p < 0.001$ in the chi2 test. Therefore, there is a possible association between POD and MetS. The odds ratio indicates that patients with POD are 2.43 times more likely to have MetS.

6.3.2 | Tradictional MetS prediction without biochemical variables

We found in the review several articles such as Murguia-Romero (Murguía-Romero et al. (2013)), Ivanovic (Ivanović et al. (2016)), and Chen (Chen et al. (2014)) using ANN to predict MetS without biochemical variables. So,we conducted several experiments to compare the models described by the authors in those articles (Chen et al. (2014); Ivanović et al. (2016); Murguía-Romero et al. (2013)). We used the data from the study of 615 subjects from the Universidad del Norte. Table 6.2 shows the variables used to build the ANN of each article.

We analyzed the training and test distributions of each of the following articles in chronological order, as published by Murguia-Romero ((Murguía-Romero et al., 2013)), Chen (Chen et al. (2014)), and Kupusinac (Ivanović et al. (2016)). For example, Murguia-Romero ((Murguía-Romero et al., 2013)) and Chen (Chen et al. (2014)) used 70/30, and Kupusinac (Ivanović et al. (2016)) used 80/10/10, as explained in the review section. We homogenized and compared all the experiments using a feed-forward Artificial Neural Network (ANN) with back-propagation of 3 layers perceptron and with the training data (70% of the data) and the testing data (30% of the remaining data). For the validation, we used the random subsampling technique of 100 times.

We implemented an ANN with 25 hidden neurons as published by (Murguía-Romero et al. (2013)). It is important to note that Murguia-Romero (Murguía-Romero et al. (2013)) did not publish all the configuration parameters of the ANN, only the number of hidden neurons, and so, we used the configuration set by Kupusinac (Ivanović et al. (2016)) using the parameters shown in Table 6.8 to build the ANN because they were the only one who published the configuration.

Table 6.8: Parameters of the ANN(Data from (Ivanović et al. (2016))).

Parameter	Value
Training Function	Levenberg–Marquardt back-propagation
min_grad	10^{-10}
mu	10^{-3}
mu_dec	0.1
mu_inc	10
mu_max	10^{10}
HL function	hyperbolic tangent sigmoid
Out function	Log-sigmoid

Murguia-Romero (Murguía-Romero et al. (2013)) used the variables WC, Sex, Height, Weight, and BMI. The results for this ANN show sensitivity 69.44%, specificity 63.78%, and AROC 74.8% with random subsampling validation of 100 times and a ratio of 70% for training data and 30% for testing data due to the low prevalence of MetS. On the other hand, Chen (Chen et al. (2014)) used the variables SEX, AGE, BMI, WC, HC, WHR, SBP, and DBP and an ANN of 5 hidden neurons. We implemented and tested it using the same configuration to homogenize and compare, resulting in sensitivity 75.37%, specificity 72.54%, and AROC 81.75% using HMS criteria.

We built the ANN published by Kupusinac (Ivanović et al. (2016)) but we changed the distribution of training and testing data (70% for distribution and 30% for testing). We used the random subsampling validation of 100 times, obtaining a mean of the sensitivity of 71.62%, a specificity of 66.95%, and AROC of 74.94%, for ANN with 96 hidden neurons. Moreover, using the ANN of 85 hidden neurons, we obtained a mean sensitivity 72.22%, specificity 66.25%, and AROC 74.79%.

In this chapter, we used an algorithm of sequential feature selection by Matlab (Guyon and Ellisseff Andre (2003); Rückstieß et al. (2011)) with 17 variables from the set of variables detailed in Table 6.2 obtaining a set of AGE, WC, WHR, and SBP variables to achieve the maximum discrimination in the classification algorithms. The number of hidden neurons was calculated with Equation (6.7), resulting in 4 with the same configuration parameters by Kupusinac (Ivanović et al. (2016)). We used random subsampling validation, obtaining the performance indicators of sensitivity 66.92%, specificity 80.57%, and AROC 82.48% using HMS criteria.

As a summary, Figure 6.2 shows the performance indicators of each experiment to compare with the other three models of ANN differentiating only in the number of hidden neurons and the input variables.

The behavior of the data mining techniques shows in Figure 6.2 that the ANN of

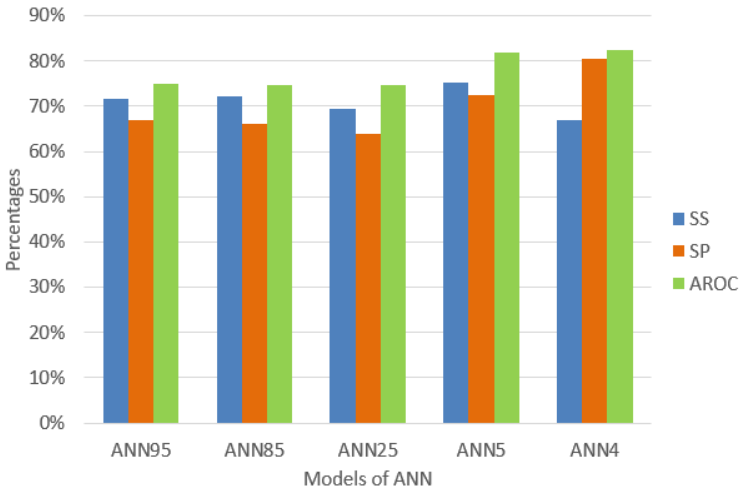

Figure 6.2: Percentage of the performance indicators of the models of ANN.

4 hidden neurons is better compared to the previously proposed techniques. It appears that decreasing hidden neurons increases AROC, and this is a reason for some researchers to estimate hidden neurons empirically until the minimum number of neurons required is found. However, we decided to use a common method to calculate it (Boger and Guterman (1997); Karsoliya (2012); Panchal and Panchal (2014)) and had obtained promising results. The diagnostic of traditional MetS without doing a blood test presents an excellent level to discriminate (AROC = 82.48%) using only four (4) variables AGE, WC, WHR, and SBP using an ANN of 4 hidden neurons. We found significant AROC values to determine an excellent MetS classification without using a blood sample and, at the same time, knowing which anthropocentric and clinical variables caused it.

However, based on clinical trials performed by the National Heart, Lung, and Blood Institute (NHLBI), excellent management of the individual risk factors of the syndrome should prevent or delay the onset of diabetes mellitus, hypertension, and cardiovascular disease (Alberti et al. (2009); Alshehri (2010); Duncan et al. (2003)).

MetS is a combination of five risk factors. For a MetS diagnosis, it is necessary to calculate the risk factors' values according to the decision threshold shown in Table 6.1. An example case of MetS implies the combinations of the dichotomous variables represented by the format $(W \times P \times G \times H \times T)$ in base 2. This example case diagnoses MetS due to normal blood pressure, increased waist circumference, triglycerides, fasting plasma glucose and decreased HDL-C (W = 1, P = 0, G = 0, H = 1, and T = 1). Hence, the prevalence of W, P, G, H, and T in a dataset is significant for balancing each type of

MetS. Therefore, we analyzed the percentage of W, P, G, H, and T, and the result was 72.68%, 27.97%, 13.33%, 63.58%, and 42.76%, respectively, as shown in Figure 6.3.

6.3.3 | MetS Types prediction without biochemical variables

The experiments were performed based on the segmentation of MetS using the HMS criterion represented in Equation (6.6) that shows the ten (10) MetS types. We obtained these types by making an AND operation among each dichotomous risk factors W, P, G, H, and T using a dataset of 615 subjects, resulting in a distribution of the different (10) types of MetS shown in Figure 6.4 that shows the prevalence of each MetS type: WPT, WPH, WPG, WGT, WGH, WTH, PGH, PGT, PTH, and GHT were 14.31%, 16.75%, 5.53%, 7.32%, 9.11%, 26.18%, 3.9%, 3.9%, 10.41%, and 6.67%, respectively. These types are a built-in set giving the traditional MetS a prevalence rate of 42.60%.

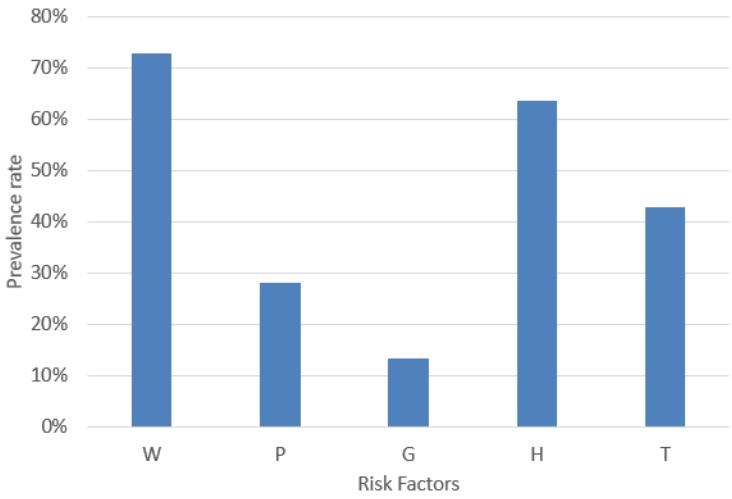

Figure 6.3: Prevalence rate of the MetS risk factors.

We can observe several MetS types with a low prevalence rate of less than 10%. We can note the risk factor (G) of fasting plasma glucose with the lowest rate (13.33%), as shown in Figure 6.3. This situation could lead to lower accuracy or AROC of prediction by the classifiers. The goal of the chapter is to classify each type of MetS. However, this research's dataset is highly imbalanced, as depicted in Figure 6.4.

Therefore, we analyzed four approaches for improving the accuracy or AROC for the different MetS types due to an imbalance of the dataset.

- Approach 1 is using only the ANN technique with a feature selection algorithm.

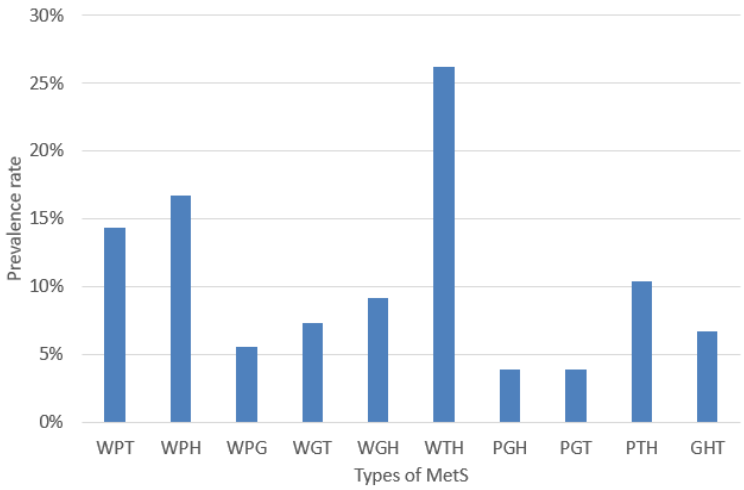

Figure 6.4: Prevalence rate of the MetS types.

- Approach 2 uses an ensemble classification algorithm in the dataset, which is the Random undersampling Boosted tree (RusBoost) ensemble.
- Approach 3 uses SMOTE to create more data that we called dataset with oversampling for then applying ANN.
- Approach 4 is using the dataset with oversampling and RusBoost.

For the approaches 1 and 3, we used for each MetS type a feed-forward Artificial Neural Network (ANN) with back-propagation of 3 layers perceptrons and with the training data (70% of the data) and the testing data (30% of the remaining data). For the validation, we used the random subsampling technique of 100 times.

It should be noted that approaches 1 and 2 used the original dataset of 615 samples, split into training and testing groups. However, approaches 3 and 4 used a dataset of 799 samples obtained using smote to create synthetic data and then splitting training or testing data.

6.3.3.1 | Approach 1: Prediction of Each MetS Type Using the Original Dataset and ANN

We did the following to predict each MetS type without a blood test. We first selected the necessary features to achieve the maximum discrimination in the classification algorithms using a sequential feature selection algorithm in Matlab (Guyon and Ellisseff Andre (2003); Rückstieß et al. (2011)) using 14 variables. We obtained the set of variables detailed in Table 6.9. To compare the traditional or general MetS (MetSG) with the

MetS types, we also show the features selected to build a model to predict it without a blood sample.

Table 6.9: Selection of predicting variables for each MetS type from original dataset.

Types	**Predicting Variables**				
WPT	WC	SBP	DBP	POD	
WPH	BMI	BFP	HC	SBP	DBP
WPG	AGE	HC		SBP	
WGT		WC			
WGH		BFP		HC	
WTH		BFP		WC	
PGH		SBP		POD	
PGT	AGE	WG		SBP	
PTH	HC	SBP		DBP	
GHT		WSR			
MetSG	AGE	WC	WHR	SBP	

Table 6.9 shows the predictor variables for the MetS types. It is important to highlight the BFP variable's relationship with the MetS types WPH, WGH, and WTH. This situation occurs due to the risk factor H (dichotomous HDL-C) dependence on gender. The MetS types related to P (dichotomous Blood Pressure) such as WPT, WPH, WPG, PGH, PGT, and PTH, according to the selection algorithm, have an evident relationship with the variables SBP and DBP. On the other hand, the MetS types WGT and GHT have the predictor variables WC and WSR, respectively, presenting a great challenge to discriminate them.

We then designed each classification model for the MetS, taking into account which variables would be treated. These selected features were used as inputs of the ANN with several hidden neurons calculated using Equation (6.7) as shown in Table 6.10 and with the same configuration parameter recommended by Kupusinac (Ivanović et al. (2016)) to predict the following MetS types: WPG, WPH, WPT, WGT, WGH, WTH, PGT, PGH, PTH, and GHT according to the HMS criterion.

Table 6.10: Numbers of hidden neurons from each ANN of the MetS types.

WPT	WPH	WPG	WGT	WGH	WTH	PGH	PGT	PTH	GHT	MetSG
4	4	3	2	2	2	2	3	3	2	4

We validated each ANN using random subsampling, which we explained previously, obtaining the average performance indicators as shown Figure 6.5. The classification algorithms' performance indicators for diagnosing the MetS type WPT show an outstanding ability to discriminate (AROC = 90.58%). Also, the prediction of the MetS type WPH has the same level of outstanding discrimination ability (AROC = 92.85%). The MetS type WPG prediction shows an excellent level to discriminate (AROC = 85.28%). The PGT and the PTH types also show an excellent level due to (AROC = 81.06%) and (AROC = 88.84%), respectively. The MetS type WGH, WTH, and PGH show an acceptable level (AROC =71.93%), (AROC =70.60%), and (AROC =73.03%) respectively.

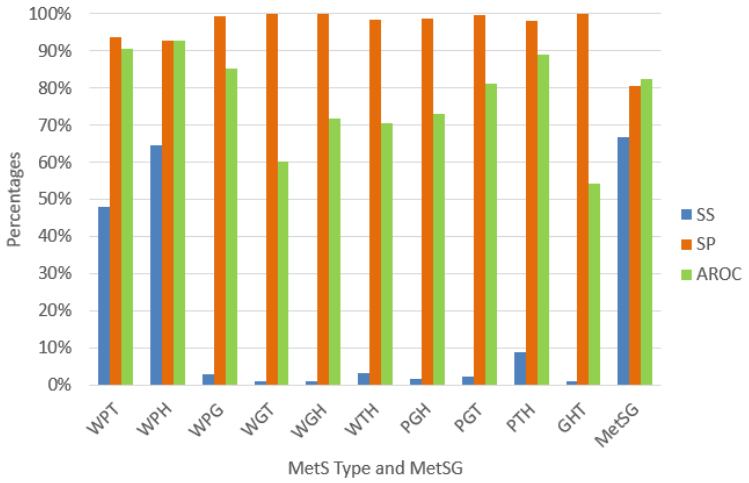

Figure 6.5: Performance indicators of the ANN for the MetS types using the original dataset.

In contrast, for the prediction of the MetS types WGT, GHT, results show a substandard level of ability to discriminate (AROC = 60.13%), and (AROC = 54.13%) respectively. This situation is due to the low levels of sensitivity. A reason for this could be that there are so few predictive variables or low prevalence rates for each type, as shown in Figure 6.4. The figure shows that most of the MetS types have a prevalence rate of less than 10% and are affected or generated by the less prevalence rate of fasting plasma glucose of 13.33% compared to the other risk factors.

6.3.3.2 | Approach 2: Prediction of Each MetS Type Using the Original Dataset and RusBoost

The approach 2 is to use the same variables' selections of the dataset and an ensemble classification algorithm. The selected algorithm is the ensemble Random undersampling Boosted tree (RusBoost), which we explained previously. This classifier is appropriate for imbalanced data. We run RusBoost with the variables selections of Table 6.9 and with the configuration described by Table 6.5 to validate with random subsampling obtaining the average performance indicators given in Figure 6.6 that shows interestingly the RusBoost technique obtained excellent levels for the AROC performance indicator in the prediction of the MetS types WPT, WPH, WPG, PGH, PGT, and PTH. The values were 88.56%, 89.79%, 83.67%, 81.04%, 81.30%, and 83.33% respectively with improvements in sensitivity rate.

On the other hand, the results show lower AROC levels for the MetS types WGT, WGH, WTH, and GHT, with values of 66.11%, 62.58%, 64.71%, and 51.53%, respectively. For the MetS types WGT and GHT, their few predictors variables possibly affected the performance indicators. The same happened with other MetS types with a low prevalence of fasting plasma glucose that created this effect, generating an imbalance in the rest of the other data and affecting the AROC levels.

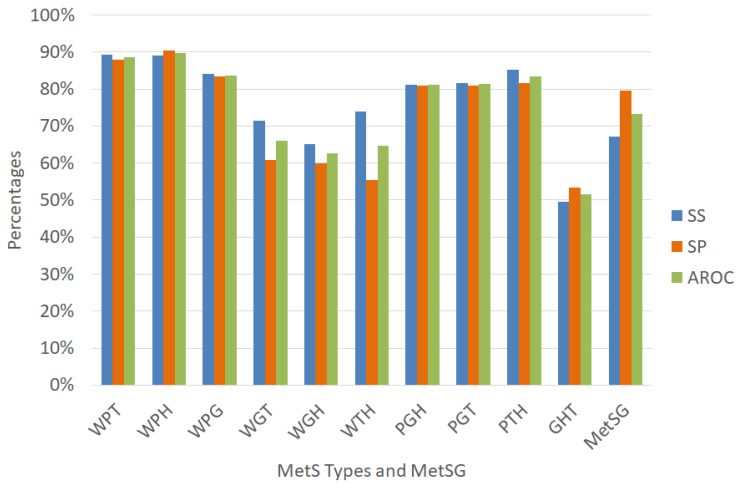

Figure 6.6: Performance indicators of the RusBoost for the MetS types using the original dataset.

6.3.3.3 | Approach 3: Prediction of Each MetS Type Using the Dataset with Oversampling and ANN

The approach 3 is to solve the imbalanced dataset for each MetS type. Several research articles have proposed that the balanced dataset improves prediction (Bouwmeester et al. (2012); Sun et al. (2009)), and most importantly, it improves ANN training as the model can correctly adapt to the minority feature of the data. Therefore, we used the sampling methods technique to balance the dataset to improve the representation of each MetS type (Melillo et al. (2013)). For this reason, a data balancing algorithm called SMOTE (Chawla et al. (2002); Fernandez et al. (2018)) was used with the WEKA data mining tool. In the dataset of 615 patients, the fasting plasma glucose dichotomous variable (G) was used in SMOTE to generate synthetic data (Bolón-Canedo et al. (2013)) with a result of 799 samples (615 plus 184 synthetic data). This approach increased the prevalence of MetS to 51.81%, and improved the prevalence rate of fasting plasma glucose. It also updated the distribution of the W, P, G, H, and T risk factors. The new values were 73.72%, 29.16%, 30.79%, 65.58%, and 45.93%, respectively, as observed in Figure 6.7. We called the new dataset of 799 samples as dataset with oversampling.

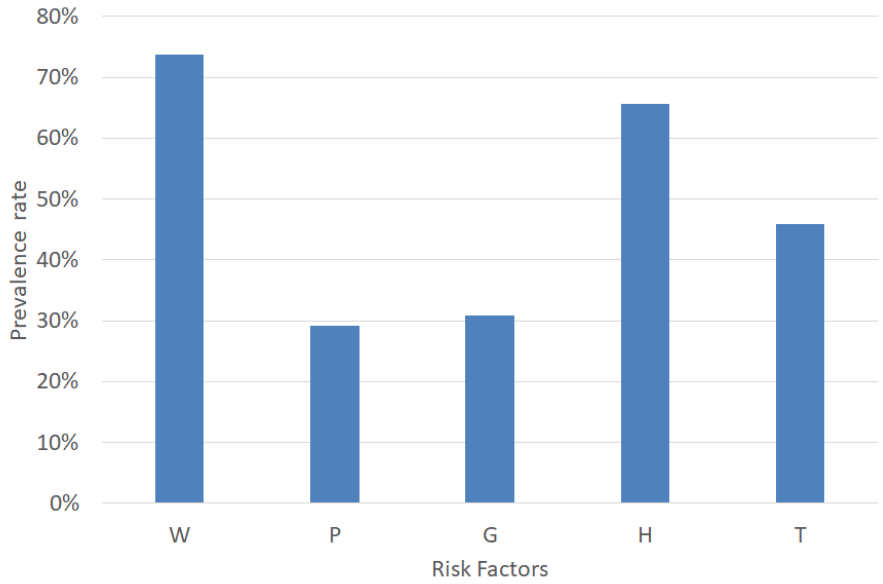

Figure 6.7: Prevalence rate of the MetS risk factors using the dataset with oversampling.

Moreover, as a result of using SMOTE in the fasting plasma glucose (G) risk factor, the percentage of the MetS types related to fasting plasma glucose increased, as shown in Figure 6.8. This result shows that the prevalence rate of the MetS types WPG, WGT, and WGH is greater than 10%, PGH and PGT is greater than 5%, and GHT s higher

than 15% in comparison with Figure 6.4. The prevalence rate of WPT, WPH, WPG, WGT, WGH, WTH, PGH, PGT, PTH, and GHT was 15.39%, 18.15%, 11.89%, 17.52%, 22.9%, 29.41%, 9.01%, 8.01%, 11.14%, and 15.39%, respectively. This result generated an increment in the prevalence rate of the traditional MetS of 51.81% as well.

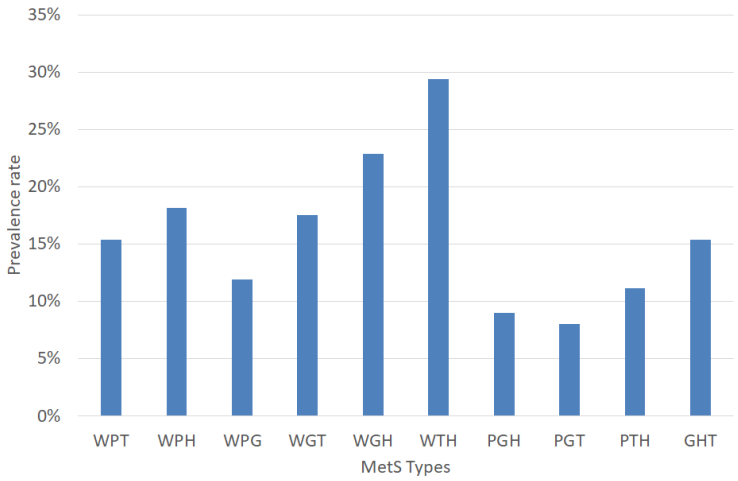

Figure 6.8: Prevalence rate of the MetS types using the dataset with oversampling.

We used an algorithm of sequential feature selection from Matlab to achieve maximum discrimination in the classification algorithms, and Table 6.11 shows the results. This step refined the variables' selection due to synthetic data's creation, similar to the actual data. This step increments the positive values of the MetS types related to biochemical variables, especially those related to fasting plasma glucose. As part of this refinement, the types WPG and WGT selected POD as a predictor variable. On the other hand, the BFP variable remains in both Tables 6.9 and 6.11, especially for the types WPH, WGH, and WTH. In those types, the biochemical variable HDL-C is related, and this depends on gender. The BFP variable is a function of gender, waist circumference, and age, demonstrating a logical relationship with WPH, WGH, and WTH types. Interestingly, the WPG and PGH types have the same initial predictor variables.

Afterward, we designed several ANN for each type of MetS according to the variables of Table 6.11 using the dataset with oversampling. These selected features were used as inputs of the ANN with several hidden neurons calculated using Equation (6.7) as shown in Table 6.12 and with the same configuration parameter recommended by Kupusinac (Ivanović et al. (2016)) and was validated each ANN using random subsampling obtaining the average performance indicators as shown Figure 6.9.

Figure 6.9 shows the ANN to classify for the prediction of the MetS types WPT,

Table 6.11: Selection of predicting variables for each target from the dataset with oversampling.

Target	Predicting Variables				
WPT	WC	SBP		DBP	
WPH	BFP	HG	WHR	SBP	DBP
WPG	AGE	POD	WG	SBP	
WGT		WC		POD	
WGH	BFP	HC	WHR	DBP	
WTH		BFP		WC	
PGH	AGE	POD	WG	SBP	
PGT	AGE		WG	SBP	
PTH	BFP	HC	SBP	DBP	
GHT		HC		POD	
MetG	AGE	WC	WHR	SBP	

Table 6.12: Numbers of hidden neurons from each ANN of the MetS types.

WPT	WPH	WPG	WGT	WGH	WTH	PGH	PGT	PTH	GHT	MetSG
3	4	4	2	4	2	4	3	4	2	4

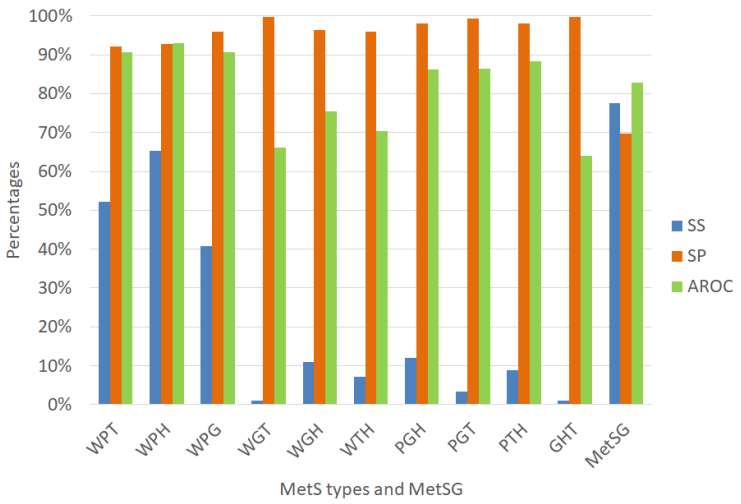

Figure 6.9: Performance indicators of the ANN for the MetS types using the dataset with oversampling.

WPH, WPG were an outstanding ability to discriminate given that the AROC were 90.69%, 93.06%, and 90.57%, respectively with an excellent specificity rate. The MetS types PGH and PTH showed an excellent ability to discriminate AROC of 86.32% and 88.41%, respectively, with an excellent specificity rate. The PGT, WGH, and WTH types showed an acceptable level (AROC = 76.92%), (AROC = 75.52%), and (AROC = 70.38%) respectively, with a regular sensitivity rate.

In contrast, the prediction of the MetS types WGT and GHT showed a regular level (AROC = 66.22%) and (AROC = 64.06%), respectively. Moreover, the sensitivity rate was almost null due to the few positive cases that have that MetS type, generating overfitting in the ANN in the training stage, since it has learned more negative cases than positive ones. Therefore, these two models are not reliable in these conditions.

The prevalence of WGT and GHT is similar to the other MetS types that have an excellent level of AROC, such as WPT, WPH, WPG. Therefore, we think that the prevalence in this dataset with oversampling is not reason. However, the traditional MetS prediction improved its AROC level to 82.86% (Excellent).

6.3.3.4 | Approach 4: Prediction of Each MetS Type Using the Dataset with Oversampling and RusBoost

Approach 4 uses the same variables selection from the dataset with oversampling of Table 6.11 and the ensemble Random undersampling Boosted tree (RusBoost) algorithm using the configuration described in Table 6.5 in the Methodology section to validate using random subsampling obtaining the average performance indicators given in Figure 6.10.

Figure 6.10 shows that the prediction accuracy values of the MetS types increased when compared with the results shown in Figure 6.6 and also the AROC discrimination ability values especially to predict the MetS type WGH due to obesity, high fasting plasma glucose, and low HDL-C with an acceptable level to discriminate (AROC = 71.08%) which increased 8.5%. The MetS type WPH showed an outstanding ability to discriminate with an AROC of 91.49% with an excellent sensitivity rate. The MetS types WPT, WPG, PGH, PGT, and PTH, showed an excellent ability to discriminate AROC 89.20%, 85.36%, 84.20%, and 84.10% respectively, with very good sensitivity rate. The MetS types WGH and PGT showed an acceptable level (AROC = 71.08%) and (AROC = 78.39%), respectively, with an acceptable sensitivity rate.

In contrast, the MetS types WGT, WTH, and GHT showed an regular level (AROC = 65.82%), (AROC = 65.16%), and (AROC = 65.65%), respectively. The prevalence of WGT, WTH, and GHT is similar to the other MetS types with an excellent level of AROC such

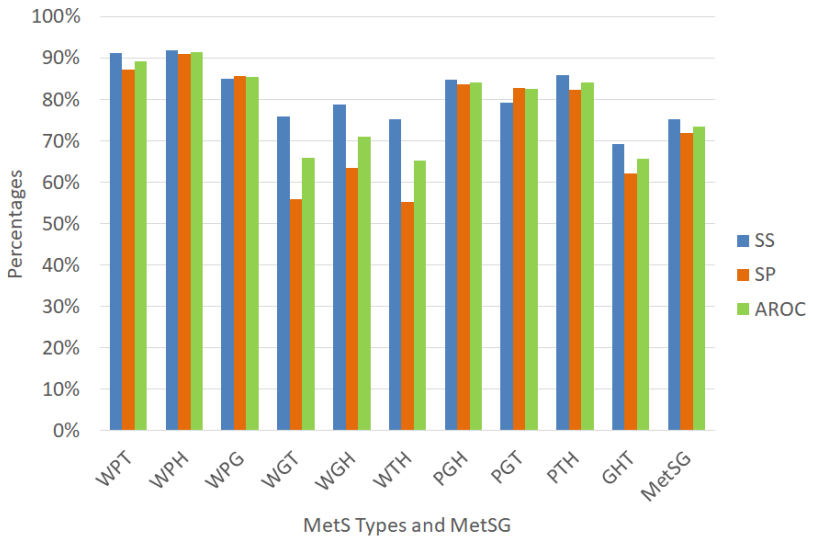

Figure 6.10: Performance indicators of the RusBoost for the MetS types using the dataset with oversampling.

as WPT, WPH, WPG.

6.4 | Discussion

One important issue to discuss is the AROC low level for some of the MetS types. The AROC low level in the WGT, WTH, and GHT types is not related to the prevalence of each MetS type in the balanced dataset. However, the reason could be the few predictors variables of the models. On the other hand, it is interesting to note that when SMOTE is used to balance a risk factor such as fast plasma glucose in a dataset for traditional MetS prediction, minimal improvement in AROC levels is achieved. For ANN, the difference is 73.50% compared to 73.25%. For RusBoost, the difference is 82.48% compared to 82.86% using the same variables AGE, WC, WHR, and SBP.

Each medical organization established its criteria to predict the MetS, which varies according to the thresholds of risk factors. All the cited medical organizations have in common that for the prediction of the MetS, doctors must check at least three of five risk factors. Therefore, it is a combination of five risk factors, in which at least three are positive.

This chapter is an example of the conjunction of data science with the combinatorial analysis and the simplification applications by the Quine–McCluskey algorithm for finding the segmentation of the MetS. This approach achieved the prediction of the MetS

types without using a blood sample that is, using a non-invasive method. For example, for a patient with waist circumference, triglycerides, and increased blood pressure, the system would predict only one type of MetS (WPT) to be active and the other types to be inactive. Therefore, doctors can infer that the patient has MetS due to increased triglycerides, blood pressure, and waist circumference to help focus on initial treatment and prevent diabetes mellitus or stroke.

We found that each MetS type's predictor variables are different from those of the traditional MetS. These variables were used to configure each machine learning technique to predict each MetS type without a blood test. We also found that the prevalence of the MetS type related to the risk factor of fasting plasma glucose has a low rate. Therefore, we performed four approaches to improve the performance indicators of the classifiers.

The first approach was to use artificial neural networks to predict each of the MetS types working with previously selected variables, obtaining excellent AROC levels for the types of MetS such as WPG, WPT, WPH, WGH, WTH, PGT, PGH, and PTH. However, the sensitivity levels were low when diagnosing some MetS types indicating a high type 2 error level. The second approach was to use an algorithm specialized in data imbalance to compensate for the sensitivity levels and the specificity levels, thereby decreasing the ROC levels affecting four types of MetS WGT, WGH, WTH, and GHT.

The third approach was to increase the sample with an oversampling algorithm such as SMOTE, allowing evaluation of the ANN models with selected variables from these new data, finding an increase in AROC levels. However, the sensitivity levels are relatively low in some types, and so, the type 2 errors decreased relatively but are still high. The fourth approach was a mixture of using an imbalanced data prediction algorithm RusBoost with the larger dataset. We found a significant improvement in the levels of AROC for the MetS types levels and, at the same time, a considerable increase in sensitivity and, therefore, decreased the type 2 error. This result favored its choice compared to the neural networks for the types WPT, WPH, WPG, WGH, PGH, PGT, PTH. This approach resulted in 7 types of MetS that can be diagnosed without using a blood test. However, the types WGH, WTH, and GHT have a regular level to predict it, possibly due to the few predictors variables that can be reflected in its power of discrimination.

Another interesting point is that the MetS types WPT, WPH, WPG, PGH, PGT, PTH have better performance in the AROC than traditional MetS diagnosed using anthropometric variables using ANN or RusBoost.

The result of this chapter demonstrates the existence of ten (10) types of MetS according to the HMS criteria and their diagnostic using non-biochemical variables, such as the anthropometric and clinical variables using the ANN and RusBoost. Moreover,

it demonstrates that doctors can diagnose traditional MetS using non-biochemical variables with classifiers. The results can vary according to the prevalence of the MetS types present in the dataset.

6.5 | Summary

Healthcare professionals diagnose the metabolic syndrome through 5 factors, two of which they get in a medical consultation: Waist Circumference level (W) and blood pressure level (P). However, Triglyceride, HDL-C, and fasting plasma glucose levels (T, H, G) require a blood test. When we analyzed the segmentation of the MetS in types, we observed that for the prediction, an algorithm requires three risk factors, and we proved which risk factors generate that disease. Therefore, we suggest that in the future, MetS studies should take them into account to know which MetS type a patient has.

The present work uses information that doctors can collect from medical history and the medical visit. Such data includes Previous Obesity Diagnosis (POD), Age, Height, Weight (WG), Waist Circumference (WC), Hip Circumference (HC), Systolic and Diastolic Blood Pressure (SBP, DBP), Body Fat Percentage (BFP), and Body Mass Index. We used an algorithm of sequential feature selection and compared machine learning techniques such as ANN and RusBoost to predict the several types of MetS without doing a blood test. This discovery helps in an early screening of one or several MetS types through anthropometric and clinical data using non-invasive methods. From this point, doctors can take relevant actions to change them through habits modification.

This chapter documents the realization of four approaches to obtain the best results. The first approach was carried out using clinical data from 615 subjects of selected variables to evaluate the ANN, obtaining excellent levels of AROC in the WPG, WPH, WPT, WGH, WTH, PGT, PGH, and PTH MetS types. However, the sensitivity levels were regular, presenting a considerable rate of type 2 errors due to the data imbalance. The second approach used a classifier for imbalanced data such as RusBoost, which improved the sensitivity levels to predict each MetS type, decreasing the type 2 error rate. However, the regular AROC levels decreased particularly for the classifiers for the WGT, WGH, WTH, and GHT types.

The third approach was to use the SMOTE technique to balance the data, and in this way, we achieved an improved performance of the ANN classifiers. However, in some classifiers, the sensitivity levels were regularly presenting a considerable rate of type 2 errors. The fourth approach was to use the balanced data and the RusBoost technique. This approach generated for the MetS types WPT, WPH, WPG, WGH, PGH, PGT,

and PTH the following excellent AROC levels: 89.20%, 91.49%, 85.36%, 71.08%, 84.20%, 78.39%, and 84.10%, respectively, and with high sensitivity rates. The fourth approach obtained the best results for most MetS types, but for classifying the traditional MetS, the third approach was the best.

7

Scrum Thinking: A Framework for the Development of mHealth

The study and development of software systems to predict cardiovascular diseases have increased in the last decades. The development of Information, Communication, and Technology (ICT) facilitated new healthcare software tools, commonly known as Mobile Health applications (or mHealth). Such tools allow healthcare professionals to closely monitor patients, facilitating information exchange to improve its users' health status. This chapter proposes a SCRUM Thinking framework for creating mHealth applications that contribute to the patient's improvement. Software developers can use this framework to design and implement applications with different software components, some of which could include the prediction algorithms presented in this thesis.

The standard and fundamental characteristics of such a framework are analyzed in this chapter, as long as the quality attributes, which were verified in the early stages of SCRUM Thinking (Empathize, Define, Devise). A case study of SCRUM Thinking was also carried out when developing a mHealth to predict the metabolic syndrome without using biochemical variables, allowing continuous monitoring of the syndrome without latency.

7.1 | Introduction

There has been in the last decades in the world an alarming increase in cardiovascular risk factors. Such factors include diabetes mellitus (DM) (Isomaa et al. (2001)), high blood pressure (Lefèbvre (2003)) and obesity (Mokdad et al. (2003)). Some of these factors make up the metabolic syndrome. According to (Kahn et al. (2005)), MetS is the trigger for cardiovascular disease, which is the principal cause of death and disability

in the world (Levenson et al. (2002)). This situation is projected as a major challenge for innovative and efficient solutions (Massot et al. (2012); United Nations, Department of Economic and Social Affairs (2015)).

The proposed strategy is to prevent and promote well-being through the use of available technologies, especially the adoption of solutions based on mHealth. Previous research demonstrated the effectiveness of this technology in self-management of Type 2 diabetes Castelnuovo (2011) and obesity (Oh et al. (2015)).

Mobile Health is a set of mobile applications in the health sector that doctors and patients use for pre-diagnosis, follow-up, control, and monitoring of their health status. However, the use of proven mHealth technologies for disease prevention is barely growing since there are many to classify and forecast diseases, but few improve the smartphone user's health conditions. Therefore, when developing a mHealth that improves health, it is necessary to use a particular framework to obtain the quality attributes required to make significant changes in the end-user, that is, the patient.

Therefore, a framework is proposed, with which developers can create mHealth applications to help patients improve their lifestyles and behavior towards their health condition. The framework uses design thinking to obtain the desirable aspects to generate a paradigm shift from the patient's point of view or end-user. For example purposes, a mHealth application will be developed to estimate the metabolic syndrome risk. This framework then builds on top of the previous chapters, given that the idea is to complement the prediction of the syndrome with a framework to design mobile applications that can help patients improve their lifestyle supported by the predictions the tool can give to the healthcare professionals.

7.2 | Background

The search for efficient and more covered health services has promoted ICT tools in the market with wide adoption or rapid growth. This approach has allowed the evolution of concepts like eHealth. Electronic health (eHealth) is the care and transfer of health resources by electronic means (as defined by WHO). An essential component of eHealth is mHealth, which is the practice of medicine and public health supported by mobile devices such as mobile phones, patient monitoring devices, personal digital assistants, and other devices(World Health Organization (2011)).

mHealth has become a promising means of promoting behavioral change for the patient's health. Using a smartphone as a platform for health interventions may be a viable way forward (Devries et al. (2013); Joe and Demiris (2013); Stephens and Allen (2013);

Yan et al. (2015)) to improve health. The mHealth are applications aimed at health professionals (either for direct use in consultation, as tools to aid diagnosis or treatment, or as tools for updating or acquiring new knowledge) or patients (self-monitoring).

The problems in the health area are also related to the management and logistics of patient care as it was presented in the Kaiser hospital networks (Carlgren (2016)) and was solved with the Design thinking (DgTh) method. Tim Brown et al. (Brown (2008)) comment that DgTh is designed by and for people. It is necessary to know the end-users' tastes and desires, investigate, and observe them as input for the design and development of any product or service that adapts to their needs.

Design thinking is a five (5) stage methodology, as shown in Figure 7.1, which are described below:

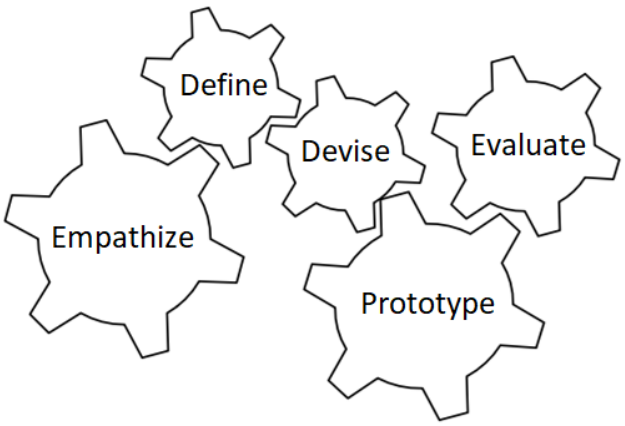

Figure 7.1: Stages of Design Thinking(Brown (2008))

- Empathize: Implies having a deep understanding of the needs of users and their environment. Developers need to put themselves in the shoes of the users to generate solutions consistent with their realities.
- Define: Implies filtering the information collected and keep what adds value and brings new perspectives within reach. Identify problems whose solutions will be key to obtaining an innovative result.
- Devise: Implies a generation of endless options. Activities encourage expansive thinking, and value judgments must be eliminated.
- Prototype: Moving from idea to reality: build prototypes, which help visualize the elements that must be improved or refined before reaching the final result.

- Evaluate: Test the prototypes with the users involved in the development of the solution. Identify significant improvements, failures to resolve, possible deficiencies. During this phase, the idea evolves until it becomes the solution the designer is searching for.

Designers need to understand the customer, their needs and desires, and positively impact them and the brands behind each solution. Based on the above, a review of the literature on mHealths that improved metabolic syndrome risk was performed by conducting tests and monitoring patients. The most used criteria for diagnosing Metabolic Syndrome is the International Diabetes Federation (IDF)(Alberti et al. (2006)), which establishes a consensus on the definition of the syndrome. The Metabolic Syndrome is a group of symptoms reflected in the risk factors detailed in Chapter 3. The first factor in meeting the criteria is obesity, measured by the waist circumference that must be above the decision threshold and then two or more risk factors to diagnose the syndrome.

However, several researchers have proposed several data mining algorithms for the prognosis of metabolic syndrome without using biochemical variables, that is, without using a blood test to monitor the patient continuously and not have latency in the results (Barrios et al. (2019); Kroon et al. (2008)).

7.3 | Literature review

A then several publications propose mHealth and its characteristics are given are presented more remarkable.

Willler et al (Willers and Hahn (2013)) conducted a study on a group of subjects who accepted a weight reduction program and took the data before and after the program to make a risk assessment using two different diagnostic tools: Metabolic syndrome and cardiovascular risk score (SCORE). It also contributes how significant a program is to improve the risk factors for metabolic syndrome. In this study, weight loss was achieved, which improved the variables of blood pressure and waist circumference.

Tani et al (Tani et al. (2012)) exposes that the main causes of Ischemic Heart Disease (IHD) are obesity, hypertension, arteriosclerosis, hyperglycemia and other metabolic disorders. These conditions are related to lifestyle problems, such as diet and exercise. Managing the diet becomes more difficult as the patient's condition worsens and that is why they focused primarily on behavioral changes.

To increase user awareness of food intake, a number of features were added to the developed system: an input of user lifestyle information , a calculation of total calorie intake, and a reference of model images of food in standard quantities of 80 kcal. The

ICD covers many of the causes of heart failure (HF). Appropriate management tools for CI are few. The main functions of our system were described and promoted self-management as a requirement for ICD and heart failure.

Jung et al (Jung et al. (2012)) achieved significant reduction in waist circumference, Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) between the beginning and 6 months of monitoring, triglycerides also decreased from 247 to 212 mg/dL and fasting glucose decreased from 110 to 103 mg/dL. The proportion of study participants who had at least three or more risk factors at the start of the study decreased at 6 months of follow-up through telephone counseling , mobile phone messages (SMS), and email messages. .

Bond et al (Bond et al. (2014)) conducted smartphone-based interventions that produced significant reductions in sedentary time among overweight and obese adults via smartphone application, phone calls, website.

As demonstrated in the literature review on the use of mHealth to improve the risk of metabolic syndrome, therefore, it is increasingly used as a tool to support monitoring and control in order to obtain positive results and not only quantify or classify a disease.

Also in the literature review, the lack of a framework when developing a mobile health application aimed at improving the patient's health and well-being conditions is evident (Stuckey et al. (2011, 2013)). Consequently, a framework for mHealth was developed which is described below Scrum Thinking.

7.4 | SCRUM Thinking

By merging the design thinking methodology with the Scrum framework to manage software development projects, a framework is proposed that hereafter called Scrum Thinking(ScTh).

Next, described in Figure 7.2 the phases supporting the proposal the framework Scrum Thinking wherein the steps of Design Thinking are articulated stages of empathy, Definition and contriving more stages of prototype and tests are borne by Desirable requirements of the patient to later generate functional increases through the framework where each of the roles are:

- **Product owner:** is the only person authorized to decide on which functionalities and functional characteristics the product will have. He is the one who represents the client, users of the software and all those parties interested in the product.

  Functions and responsibilities:

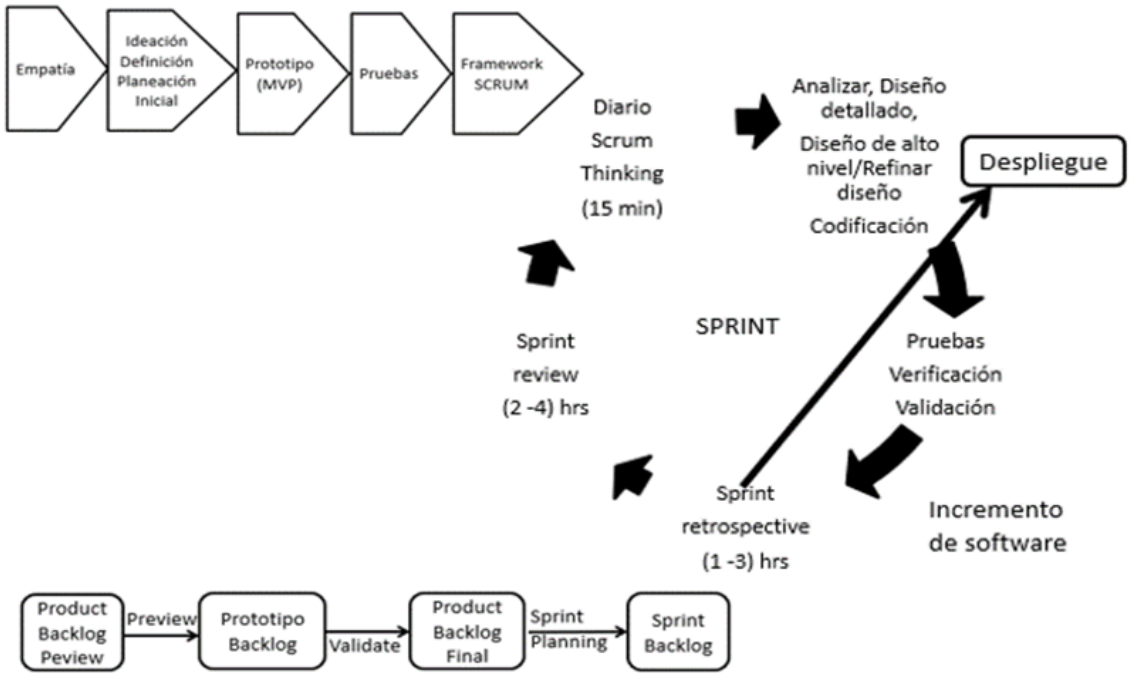

Figure 7.2: Scrum Thinking framework.

- Channel business needs , knowing how to "listen" to the parties interested in the product and transmit them in "value objectives for the product", to the Scrum Thinking team.
- Maximize business value with respect to Return on Investment (ROI), advocating for business interests.
- Review the product and go adaptándole their works l idades, analyzing the improvements that they can give a greater value to the business.

- **Scrum Thinking Master:** It is the soul of Scrum Thinking. He is not a "leader", since the Scrum Thinking Master is not a typical leader, but is a true neutral Server, who will be in charge of promoting and instructing on the agile principles of Scrum.

  Functions and responsibilities:

  - Guarantee the correct application of Scrum Thinking. This includes, from the correct transmission of its principles to top management, to the prevention of role reversals (take special care that the product owner does not act on behalf

of the Scrum Thinking Team and vice versa, or that the audience interferes in tasks that are not conducive to him).

- Resolve conflicts that hinder the progress of the project.
- Encourage and motivate the Scrum Team Thinking, creating a climate of collaborative work, it encourages self-management equipment and prevent third - party intervention in the management of the team.

- **Scrum Thinking Team:** is the multidisciplinary team of developers, made up of programmers, designers, architects, testers and others, who will be in charge of developing the product in an managed way.

  Functions and responsibilities:

  - Take the Product Backlog to potentially functional
  - Take the Product Backlog to operational developments.

  The activities of each of the roles are detailed in Tabla 7.1 , according to the phases of the framework Scrum Thinking and deliverable at the end of each phase.

7.5 | Results and Discussion

In order to verify the expected functioning of the proposed Scrum Thinking framework in the previous item, a prototype was implemented in the first iteration that allows quantifying or estimating the risk of metabolic syndrome and at the same time being understandable for the doctor and patients. The prototype, that is, Minimum Viable Product (MVP) is shown in its main view in figure 7.3.

P-MetS v01 was performed with AppInventor and is based on the tree algorithm decision to predict the metabolic syndrome from variables WCD: Waist circumference categorized according to threshold, BPD: Pressure is systolic or diastolic blood categorized by threshold, HC: hip circumference (cm) as shown in Figure 7.4.

The health professionals of the work team tested the application found excellent aspects and others for improvement. Patients recommended adding follow-up via SMS and reminders via phone calls as noted in some of the literature review projects.

Table 7.1: Scrum Thinking framework activities and products

PHASES	ROLES	ACTIVITIES	PRODUCTS
Empathy	Scrum Thinking Team Product Owner	Make observations Conduct interviews Build empathy maps Create of opportunities Create of opportunities Benchmarking	Response document Analysis documents Map, Audios, Videos
Initial planning, Definition, Ideation	Scrum Thimking Master Scrum Thinking Team	Perform the Scamper method Plan the iteration Manage changes Define concepts Analyze needs Donate criteria	Result of the Scamper method Iteration plan Documents of changes Definition of concepts Needs analysis Donate criteria
Prototype	Scrum Thimking Master Scrum Thinking Team	Define the MVP Developing the Style guide Developing the Style guide Make the visual design Build the Wireframes	MVP Style guide Visual design Wireframes
Tests	Scrum Thinking master Scrum Thinking Team Product Owner	Validate user Estimate complexity Identify risks User journeys User stories Learn lessons	Validate user Estimate complexity Identify risks User journeys User stories Learn lessons
Product backlog	Scrum Thinking master Scrum Thinking Team Product Owner	Sprint planning Define the product backlog Assign story points to user histories	Sprint plan Prioritized product backlog
Sprint Backlog	Scrum Thimking Master Scrum Thinking Team	Conduct daily meetings Computer sync	Requirements Use cases State diagram Establishment of cardiovascular risk quantification variables
Execution the sprint	Scrum Thimking Master	Running the iteration	Modeling, high-level design document Design and implementation of the solution, source code of the m-health solution
Modeling and design	Software architect ScrumThinking Master Scrum Thinking Team Stakeholders	Demonstration of completed requirements	Data model and database
Integration and deployment	Software architect Software architect Stakeholders Scrum Thinking Master Scrum Thinking Team	Do integration planning Perform verification and validation tests Run the installation Generate documentation	Mobile app Troubleshooting Manual-t

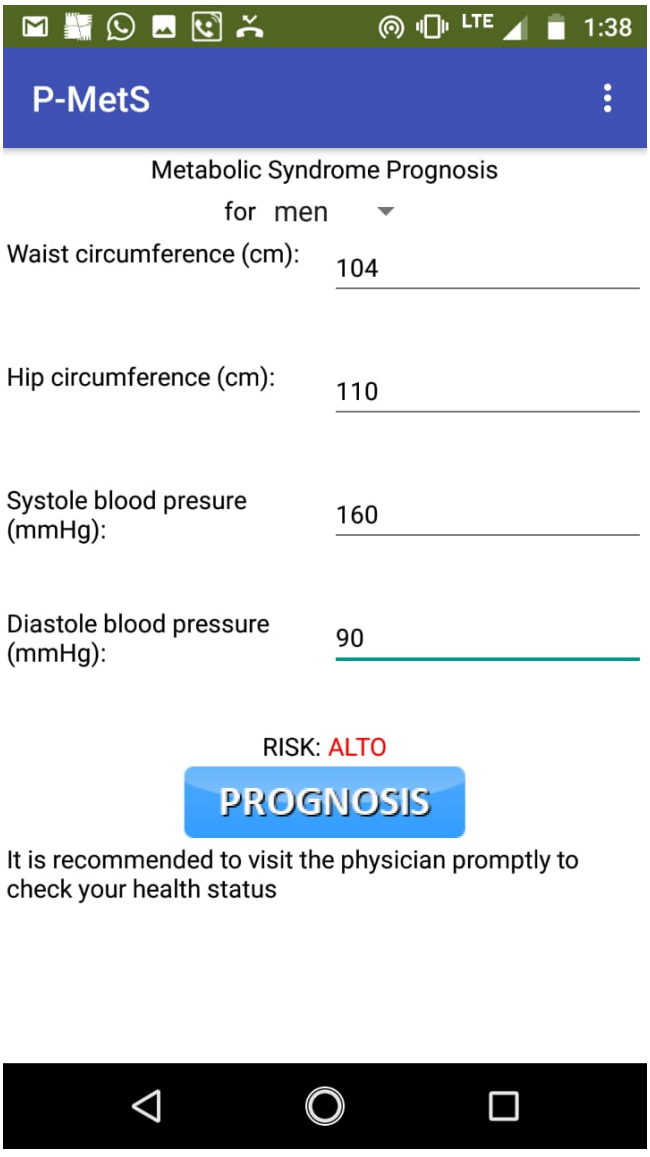

Figure 7.3: P-MetS v01.

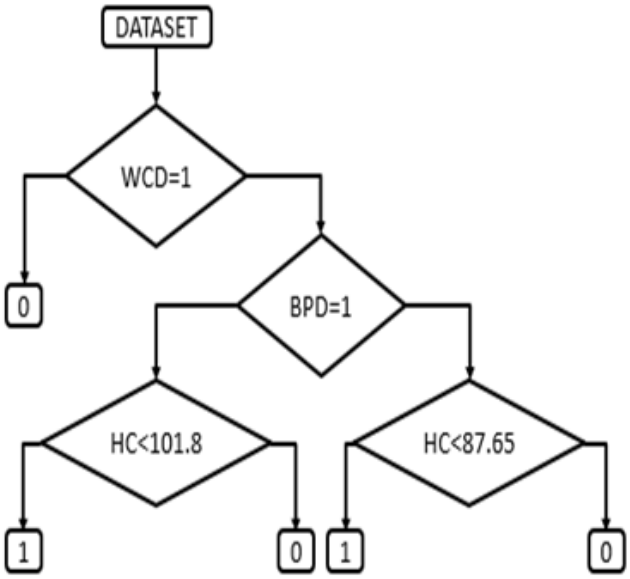

Figure 7.4: Decision Tree to predict the MetS.

In the second iteration will perform with Xamarin(Boushehrinejadmoradi et al. (2015); Vishal and Kushwaha (2018)) development of the P-MetS v02. This application will continue bulletins recommendation on the management of alarms for diet and taking medication as medical activities and appointments.

As a result of the review of previous works, it has been found that m Health is a topic with strong dynamics and a large number of previous publications. However, limiting the with those who are focused on getting effectively improved health and well-being through ICT tools such as SMS, phone calls and preset alarms, the number of investigations is greatly reduced.

The different mHealth published in the reviewed articles present an advance in the diagnosis and monitoring of the patients' health. However, the lack of a software development framework such as Scrum Thinking (ScTh), prevent it from achieving what patients really want. Scrum Thinking, due to its integration with the Design thinking method, allows obtaining in several iterations what patients really want, which, in short, is a paradigm shift in their health.

It should be noted that in the mHealth industry, a framework based on business architecture principles with close communication with stakeholders applied to the health field has not been established either.

7.6 | Summary

The relationship between health and technology has found in mobile devices a valuable tool for executing its activities that dates back a long time. However, the massification of these technological resources has also impacted the habitual health paradigm, substantially modifying the access and effectiveness of health services in prevention, prediction, location, and intelligence. The integration of Design Thinking with the Scrum framework for the development of mHealth allows a significant change in the paradigm of patient health to be obtained since in the early stages of the Scrum Thinking framework, it is ensured that both functional design requirements and quality attributes are satisfactory from the point of view of the patient or end-user, because achieving to improved health.

The proposed Scrum Thinking framework can be adapted to various applications in health technology with very few changes between them and covering a broad spectrum of smartphones. A more ambitious design cycle should include a usability testing phase with real patients to get better input from the final users and include suggested changes to the software's next versions.

8

Conclusions

This chapter summarizes the contributions of this thesis. This work analyzed the prediction models of the metabolic syndrome without a blood test based using machine learning published in the scientific community. The literature review process initially selected several publications related to this work's topic but finally discarded most of them for not being close enough. The closely related publications used different machine learning models. The models include one tree decision model, one logistic regression model, and three artificial neural networks models. However, not all share the same performance indicators or the same validation technique. Therefore this thesis proposed a methodology called RAMAD to standardize the comparison of proposed models versus existing models. RAMAD should standardize performance indicators and validation in the models to build a gold standard and compare it with a proposed model.

This thesis analyzed several models that researchers can use to determine metabolic syndrome early without doing a blood test based on machine learning techniques. The model proposed by the author to develop and validate an artificial neural network using non-biochemical variables such as anthropometric and clinical information. The validation technique of the model proposed uses random subsampling to get performance evaluation indicators to compare with the models analyzed previously using the RAMAD methodology to obtain the better resulted in the AROC indicator. This thesis used the RAMAD methodology to find the model proposed for the prediction of metabolic syndrome, but researchers can adapt it for other illnesses.

Another contribution of this work is a mathematical model representing the metabolic syndrome diagnosis according to the HMS criterion, which facilitates the under-

standing of the existence of various types of metabolic syndrome. The model added combinations without repetition of 3, 4, and 5 from 5 dichotomous variables of the risk factors to segment the metabolic syndrome and subsequently minimize in a way optimal to obtain 10 types of metabolic syndrome such as WPG, WPH, WPT, WGH, WGT, WTH, PGH, PGT, PTH, and GHT.

The thesis proposes a framework to diagnose the metabolic syndrome types without using a blood test based on Artificial Neural Networks, and Random undersampling Boosted tree using non-biochemical variables such as anthropometric and clinical information. The framework works over imbalanced and balanced datasets using the Synthetic Minority Oversampling Technique. The framework used for validation the random subsampling technique to get performance evaluation indicators between the classifiers. The results showed an excellent framework for diagnosing the 10 MetS types that have Area under Receiver Operating Characteristic (AROC) curves with a range of 71% to 93% compared with AROC 82.86% from traditional MetS.

The integration of Design Thinking with the Scrum framework for the development of mHealth allows obtaining a significant change in the paradigm of patient health. In the first stages of the Scrum Thinking technique, it ensures that both the functional design requirements and the quality attributes are satisfactory from the patient or final user's point of view. It is the case of the P-MetS application that seeks to improve the health of the patient by diagnosing the metabolic syndrome early to make behavioral changes. The proposed Scrum Thinking framework can be adapted to various healthcare technology applications with very little change between them and a broad spectrum of smartphones.

A

Appendix A: Random undersampling Boosted tree (RusBoost) ensemble.

Algorithm 1 RUSBoost Algorithm(Adapted from (Seiffert et al. (2010))).

Given: Set S of examples (x_1, y_1),...,(x_m, y_m) with minority class
Weak learner (decision tree), WeakLearn
Number of iterations, T
Desired percentage of total instances to be represented by the minority class, N

Initialise $D_1(i)$=$\frac{1}{m}$ for all i
for t **do**=1,2,...,T
    Create temporary training dataset $S_t^{'}$ with distribution $D_t^{'}$ using random undersampling
    Call WeakLearn, providing it with examples $S_t^{'}$ and their weights $D_t^{'}$.
    Get back a hypothesis h_t: XxY $\rightarrow$ [0,1].
    Calculate a pseudo-loss (for S and D_t):

$$\epsilon_t = \sum_{(i,y):y_i \neq y} D_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y))$$

    Calculate the weight update parameter:

$$\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t}$$

    Update D_t:

$$D_{t+1}(i) = D_t(i)\alpha_t^{\frac{1}{2}(1+h_t(x_i,y_i)-h_t(x_i,y:y\neq y_i))}$$

    Normalise D_{t+1}: Let $Z_t = \sum_i D_{t+1}(i)$

$$D_{t+1}(i) = \frac{D_{t+1}(i)}{Z_t}$$

end for
Output the final hypothesis:

$$H(x) = \underset{y \epsilon Y}{argmax} \sum_{t=1}^{T} h_t(x, y) log \frac{1}{\alpha_t}$$

B

Appendix B: Algorithm to diagnostic MetS with IDF criteria

Algorithm 2 Algorithm to diagnostic MetS with IDF criteria.

Hit=0;
DxMetSIDF=0;
if $((SEX == 0) and (WC \geq 80)) or ((SEX == 1) and (WC \geq 90))$ **then**
    Hit=1;
    if $(SBP \geq 130) or (DBP \geq 85)$ **then**
        Hit=Hit+1;
    end if
    if $((SEX == 0) and (HDLC < 50)) or ((SEX == 1) and (HDLC < 40))$ **then**
        Hit=Hit+1;
    end if
    if $TG \geq 150$ **then**
        Hit=Hit+1;
    end if
    if $(GLU \geq 100)$ **then**
        Hit=Hit+1;
    end if
    if $(Hit \geq 3)$ **then**
        DxMetSIDF=1;
    end if
    DxMetSIDF=0;
end if

C

Appendix C: Algorithm to diagnostic MetS with HMS criteria.

Algorithm 3 Algorithm to diagnostic MetS with HMS criteria.

Hit=0;
DxMetSHMS=0;
if $((SEX == 0) and (WC \geq 80)) or ((SEX == 1) and (WC \geq 90))$ **then**
    Hit=Hit+1;
end if
if $(SBP \geq 130) or (DBP \geq 85)$ **then**
    Hit=Hit+1;
end if
if $((SEX == 0) and (HDLC < 50)) or ((SEX == 1) and (HDLC < 40))$ **then**
    Hit=Hit+1;
end if
if $TG \geq 150$ **then**
    Hit=Hit+1;
end if
if $(GLU \geq 100)$ **then**
    Hit=Hit+1;
end if
if $(Hit \geq 3)$ **then**
    DxMetSHMS=1;
else
    DxMetSHMS=0;
end if

D

Appendix D: Algorithm to diagnostic MetS risk with ATP III criteria using the DT.

Algorithm 4 Algorithm to diagnostic MetS risk with ATP III criteria using the decision tree(Adapted from (Kroon et al. (2008))

```
1: if BMI ≥ 35 then
2:     DxMetS=1;
3: else
4:     if (BMI < 35)and(BMI ≥ 30) then
5:         if (WCD)and(BPD) then
6:             DxMetS=1;
7:         else
8:             if (WCD)or(BPD) then
9:                 DxMetS=1;
10:            else
11:                if ( WCD)and( BPD) then
12:                    DxMetS=0;
13:                end if
14:            end if
15:        end if
16:    else
17:        if (BMI < 35)and(BMI ≥ 30) then
18:            if (WCD)and(BPD) then
19:                DxMetS=1;
20:            else
21:                if (WCD)or(BPD) then
22:                    DxMetS=1;
23:                else
24:                    if ( WCD)and( BPD) then
25:                        DxMetS=0;
26:                    else
27:                        DxMetS=0;
28:                    end if
29:                end if
30:            end if
31:        end if
32:    end if
33: end if
```

E

Appendix E: Solution of Quine–McCluskey algorithm to minimize the MetS types

The Quine–McCluskey algorithm aims to minimize the logical sum of products, which are the MetS types that we will call implicants. If the MetS type is a combination of 5 variables, it is represented in the algorithm as an implicant of order 0. If it is a combination of 4 variables, it would be of order 1, and if it is a combination of 3 variables, it would be of order 2.

1. We select the implicants of order 0 as show Table E.1 that were obtained from the positive logic truth table as shown Table 6.3.
2. All those implicants of order 0, where only one variable has changed its state are grouped together. The group is obtained by eliminating the changed variable of those implicants of order 1. An example is the implicant of order 0, number 7 (W'P'GHT) and number 15 (W'PGHT), which are grouped together, resulting in W'GHT, which is of order 1.
3. Then the implicants of order 1, where only one variable has changed its state are grouped together, obtained by eliminating changed variable. For example, the implicants 7,15 (W'GHT) and 23,31 (WGHT) (both implicants of order 1) are grouped together, resulting in GHT, which is of order 2.
4. This process is carried out on all the implicants of order 0, until all implicants are minimized as shown Eq. 6.6

$$MetS_{HMS} = WPT + WPH + WPG + WGT + WGH + WTH + PGT + PGH + PHT + GHT$$

Table E.1: Implicants in the minimization of the MetS types

	IMPLICANTS				
n	Order 0*	Order 1		Order 2	
7	W'P'GHT	7,15	W'GHT	7,15,23,31	GHT
11	W'PG'HT	7,23	P'GHT	11,15,27,31	PHT
13	W'PGH'T	11,2	W'PHT	13,15,29,31	PGT
14	W'PGHT'	11,3	PG'HT	14,15,30,31	PGH
15	W'PGHT	13,2	W'PGT	19,23,27,31	WHT
19	WP'G'HT	3,29	PGH'T	21,23,29,31	WGT
21	WP'GH'T	14,2	W'PGH	22,23,30,31	WGH
22	WP'GHT'	14,30	PGHT'	25,27,30,31	WPT
23	WP'GHT	15,31	PGHT	26,27,30,31	WPH
25	WPG'H'T	19,23	WP'HT	28,29,30,31	WPG
26	WPG'HT'	19,27	WG'HT		
27	WPG'HT	21,23	WP'GT		
28	WPGH'T'	21,29	WGH'T		
29	WPGH'T	22,23	WP'GH		
30	WPGHT'	22,30	WGHT'		
31	WPGHT	23,31	WGHT		
		25,3	WPG'T		
		25,29	WPH'T		
		26,27	WPG'H		
		26,30	WPHT'		
		27,31	WPHT		
		28,29	WPGH'		
		28,30	WPGT'		
		29,31	WPGT		
		30,31	WPGH		

*The symbol apostrophe (') means that the variable is negative.

References

Aguilar, M., Bhuket, T., Torres, S., Liu, B., and Wong, R. J. (2015). Prevalence of the Metabolic Syndrome in the United States, 2003-2012. *JAMA*, 313(19):1973.

Alberti, K. G. M. M., Eckel, R. H., Grundy, S. M., Zimmet, P. Z., Cleeman, J. I., Donato, K. A., Fruchart, J.-c., James, W. P. T., Loria, C. M., and Smith, S. C. (2009). Harmonizing the Metabolic Syndrome International osis Society ; and International Association for the Study of Obesity. *Circulation*, 120:1640–1645.

Alberti, K. G. M. M., Zimmet, P., and Shaw, J. (2006). Metabolic syndrome — a new world-wide definition . A Consensus Statement from the International Diabetes Federation. *Jornal compilation*, pages 469–480.

Alberti, K. G. M. M. and Zimmet, P. Z. (1998). Definition , Diagnosis and Classification of Diabetes Mellitus and its Complications Part 1 : Diagnosis and Classification of Diabetes Mellitus Provisional Report of a WHO Consultation. pages 539–553.

Alberto, L., Martínez, C., Mercedes, D., and Hernández, F. (2010). La toma de decisiones médicas como la habilidad profesional esencial en la carrera de Medicina. *Rev. Electrónica las Ciencias Médicas en Cienfuegos*, 8(1):42–45.

Alshehri, A. (2010). Metabolic syndrome and cardiovascular risk. *Journal of Family and Community Medicine*, 17(2):73.

Andrea, T. A. and Kalayeh, H. (1991). Applications of Neural Networks m Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.*, 34(9):2824–2836.

Aschner, P. (2010). Metabolic syndrome as a risk factor for diabetes. *Expert Review of Cardiovascular Therapy*, 8(3):407–412.

Azevedo, A. and Santos, M. F. (2008). KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW. *In Proceedings of the IADIS European Conference on Data Mining 2008*, pages 182–185.

Balkau, B. and Charles, M. (1999). Comment on the provisional report from the WHO consultation. European Group for the Study of Insulin Resistance (EGIR). *Diabet Med*, 16:442–443.

Barrios, M., Jimeno, M., Villalba, P., and Navarro, E. (2019). Novel Data Mining Methodology for Healthcare Applied to a New Model to Diagnose Metabolic Syndrome without a Blood Test. *Diagnostics*, 9(4):192.

Bartlett, J. G. M. (2001). EXECUTIVE SUMMARY OF THE THIRD REPORT OF THE NATIONAL CHOLESTEROL EDUCATION PROGRAM (NCEP) EXPERT PANEL ON DETECTION , EVALUATION AND. *Infectious Diseases in Clinical Practice*, 10(5):287–288.

Berner, E. and Gong, Y. (2016). *Clinical Decision Support Systems: Theory and Practice*.

Berrar, D. (2019). Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology*, pages 542–545. Elsevier.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Boger, Z. and Guterman, H. (1997). Knowledge Extraction from Artificial Neural Networks Models Intelligent Process Control Systems Hugo Guterman. *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, pages 3030–3035.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data.

Bond, D. S., Thomas, J. G., Raynor, H. A., Moon, J., Sieling, J., Trautvetter, J., Leblond, T., and Wing, R. R. (2014). B-MOBILE - A Smartphone-Based Intervention to Reduce Sedentary Time in Overweight/Obese Individuals: A Within-Subjects Experimental Trial. *PLoS ONE*, 9(6):e100821.

Boushehrinejadmoradi, N., Ganapathy, V., Nagarakatte, S., and Iftode, L. (2015). Testing Cross-Platform Mobile App Development Frameworks (T). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 441–451. IEEE.

Boutari, C., Lefkos, P., Athyros, V. G., Karagiannis, A., and Tziomalos, K. (2018). Nonalcoholic Fatty Liver Disease vs. Nonalcoholic Steatohepatitis: Pathological and Clinical Implications. *Current Vascular Pharmacology*, 16(3):214–218.

Bouwmeester, W., Zuithoff, N. P., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., Altman, D. G., and Moons, K. G. (2012). Reporting and methods in clinical prediction research: A systematic review. *PLoS Medicine*, 9(5).

Brown, T. (2008). Design Thinking. *Harvard Business Review*, Septiembre:1–9.

Bruce, K. D. and Byrne, C. D. (2009). The metabolic syndrome: common origins of a multifactorial disorder. *Postgraduate Medical Journal*, 85(1009):614–621.

Calabria, J. and Bonilla, I. (2014). *Minería de Datos Aplicación Didáctica*. Uniautonoma.

Carlgren, L. (2016). Design Thinking in innovation , in practice : the case of Kaiser Permanente. *EURAM Conference Proceedings. European Academy of Management*.

Castelnuovo, G. (2011). TECNOB Study: Ad Interim Results of a Randomized Controlled Trial of a Multidisciplinary Telecare Intervention for Obese Patients with Type-2 Diabetes. *Clinical Practice & Epidemiology in Mental Health*, 7(1):44–50.

Chandna, D. (2014). Diagnosis of Heart Disease Using Data Mining Algorithm. *International Journal of Computer Science and Information Technologies*, 5(5):1678–1680.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(2):321–357.

Chen, H., Xiong, S., and Ren, X. (2014). Evaluating the Risk of Metabolic Syndrome Based on an Artificial Intelligence Model. *Abstr. Appl. Anal.*, 2014:1–12.

Chen, J., Muntner, P., Hamm, L. L., Jones, D. W., Batuman, V., Fonseca, V., Whelton, P. K., and He, J. (2004). The Metabolic Syndrome and Chronic Kidney Disease in U.S. Adults. *Annals of Internal Medicine*, 140(3):167.

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T., and Roccella, E. J. (2003). Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*, 42(6):1206–1252.

Cleophas, T. J. and Zwinderman, A. H. (2015). *Machine Learning in Medicine - a Complete Overview*. Springer International Publishing, Cham.

Cornier, M.-A., Dabelea, D., Hernandez, T. L., Lindstrom, R. C., Steig, A. J., Stob, N. R., Van Pelt, R. E., Wang, H., and Eckel, R. H. (2008). The Metabolic Syndrome. *Endocrine Reviews*, 29(7):777–822.

D. W. Hosmer, J. and Lemeshow, S. (2004). *Applied Logistic Regression*. John Wiley & Sons.

Devries, K. M., Kenward, M. G., and Free, C. J. (2013). Preventing Smoking Relapse Using Text Messages: Analysis of Data From the txt2stop Trial. *Nicotine & Tobacco Research*, 15(1):77–82.

Duncan, G. E., Perri, M. G., Theriaque, D. W., Hutson, A. D., Eckel, R. H., and Stacpoole, P. W. (2003). Exercise Training, Without Weight Loss, Increases Insulin Sensitivity and Postheparin Plasma Lipase Activity in Previously Sedentary Adults. *Diabetes Care*, 26(3):557–562.

Esposito, K., Chiodini, P., Colao, A., Lenzi, A., and Giugliano, D. (2012). Metabolic Syndrome and Risk of Cancer: A systematic review and meta-analysis. *Diabetes Care*, 35(11):2402–2411.

Fayyad, U., Piatetsky-shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in. 17(3):37–54.

Fernandez, A., Garcia, S., Herrera, F., and Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61:863–905.

Fliotsos, M., Zhao, D., Rao, V. N., Ndumele, C. E., Guallar, E., Burke, G. L., Vaidya, D., Delaney, J. C. A., and Michos, E. D. (2018). Body mass index from Early-, Mid-, and Older-adulthood and risk of heart failure and atherosclerotic cardiovascular disease: MESA. *Journal of the American Heart Association*, 7(22).

Floyd, T. L. (2010). *Digital Fundamentals, 10/e*. Pearson Education India.

Ford, E. S., Giles, W. H., and Dietz, W. H. (2002). Prevalence of the Metabolic Syndrome Among US Adults. *JAMA*, 287(3):356.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer-Verlag New York, second edition.

Galassi, A., Reynolds, K., and He, J. (2006). Metabolic Syndrome and Risk of Cardiovascular Disease. pages 812–819.

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., Greenland, P., Lackland, D. T., Levy, D., O'Donnell, C. J., Robinson, J. G., Schwartz, J. S., Shero, S. T.,

Smith, S. C., Sorlie, P., Stone, N. J., and Wilson, P. W. (2014). 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Journal of the American College of Cardiology*, 63(25):2935–2959.

Greenes, R. (2014). *Clinical Decision Support: The Road to Broad Adoption: Second Edition*.

Grundy, S. M. (2006). Metabolic Syndrome: Connecting and Reconciling Cardiovascular and Diabetes Worlds. *Journal of the American College of Cardiology*, 47(6):1093–1100.

Grundy, S. M. (2007). Metabolic Syndrome: A Multiplex Cardiovascular Risk Factor. *The Journal of Clinical Endocrinology & Metabolism*, 92(2):399–404.

Grundy, S. M. (2008). Metabolic Syndrome Pandemic. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 28(4):629–636.

Gutiérrez-Solis, Datta Banik, S. (2018). Prevalence of Metabolic Syndrome in Mexico: A Systematic Review and Meta-Analysis. Metabolic Syndrome and Related Disorders. *METABOLIC SYNDROME AND RELATED DISORDERS*, XX(Xx):1–11.

Guyon, I. and Ellisseff Andre (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hsiung, D. Y., Liu, C. W., Cheng, P. C., and Ma, W. F. (2015). Using non-invasive assessment methods to predict the risk of metabolic syndrome. *Appl. Nurs. Res.*, 28(2):72–77.

Irving, G., Neves, A. L., Dambha-Miller, H., Oishi, A., Tagashira, H., Verho, A., and Holden, J. (2017). International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ Open*, 7(10):e017902.

Isomaa, B., Almgren, P., Tuomi, T., Forsen, B., Lahti, K., Nissen, M., Taskinen, M.-R., and Groop, L. (2001). Cardiovascular Morbidity and Mortality Associated With the Metabolic Syndrome. *Diabetes Care*, 24(4):683–689.

Ivanović, D., Kupusinac, A., Stokić, E., Doroslovački, R., and Ivetić, D. (2016). ANN Prediction of Metabolic Syndrome: a Complex Puzzle that will be Completed. *Journal of Medical Systems*, 40(12):264.

J. Shao, J. A. (2005). Linear model selection by cross-validation. *Journal of Statistical Planning and Inference*, 128(1):231–240.

Joe, J. and Demiris, G. (2013). Older adults and mobile phones for health: A review. *Journal of Biomedical Informatics*, 46(5):947–954.

Jover, A., Corbella, E., Mun, A., Pedro-botet, J., Herna, A., and Zu, M. (2011). Prevalence of Metabolic Syndrome and its Components in Patients With Acute Coronary Syndrome. 64(7):579–586.

Jung, H., Lee, B., Lee, J.-E., Kwon, Y.-H., and Song, H. (2012). Efficacy of a programme for workers with metabolic syndrome based on an e-health system in the workplace: a pilot study. *Journal of Telemedicine and Telecare*, 18(6):339–343.

Kahn, R., Buse, J., Ferrannini, E., and Stern, M. (2005). The Metabolic Syndrome: Time for a Critical Appraisal: Joint statement from the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care*, 28(9):2289–2304.

Kaplan, N. (1989). The deadly quartet. Upper-body obesity, glucose, intolerance, hypertriglyceridemia, and hypertension. *Arch. Intern. Med*, 149:1514 – 1520.

Karsoliya, S. (2012). Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture. *International Journal of Engineering Trends and Technology*, 3(6):714–717.

Kaur, J. (2014). A Comprehensive Review on Metabolic Syndrome. *Cardiol. Res. Pract.*, 2014:21.

Kotu, V. and Deshpande, B. (2014). *Predictive Analytics and Data Mining*. Morgan Kaufmann.

Kroon, M. L. A. D., Renders, C. M., Kuipers, E. C. C., Wouwe, J. P. V., Buuren, S. V., Jonge, G. A. D., and Hirasing, R. A. (2008). Identifying metabolic syndrome without blood tests in young adults — The Terneuzen Birth Cohort. 18(6):656–660.

Kumar, S. (2012). *NEURAL NETWORKS*. Tata McGraw-Hill Education.

Lean, M. E., Han, T. S., and Deurenberg, P. (1996). Predicting body composition by densitometry from simple anthropometric measurements. *American Journal of Clinical Nutrition*, 63(1):4–14.

Lefèbvre, P. J. (2003). The metabolic syndrome revisited. *International Congress Series*, 1253:3–10.

Levenson, J., Skerrett, P., and Gaziano, J. (2002). Reducing the global burden of cardiovascular disease: the role of risk factors. *P Rev Cardiol*, 5:188–189.

Liang, Q.-S. X. and Y.-Z. (2001). Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.*, 56(1):1–11.

Lombo, B., Satizbal, C., Villalobos, C., Tique, C., and Kattah, W. (2007). Prevalencia del síndrome metabólico en pacientes diabéticos. *Acta Medica Colombiana*, 32:9–15.

Lombo, B., Villalobos, C., Tique, C., and Satizábal, C. (2006). Prevalencia del síndrome metabólico entre los pacientes que asisten al servicio clínica de hipertensión de la Fundación Santa Fe de Bogotá. *Revista Colombiana de Cardiología*, 12:472–478.

Marhl, M., Grubelnik, V., Magdič, M., and Markovič, R. (2020). Diabetes and metabolic syndrome as risk factors for COVID-19. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):671–677.

Marsland, S. (2014). *Machine Learning: An Algorithmic Perspective, Second Edition*. Chapman & Hall/CRC, 2nd edition.

Massot, B., Baltenneck, N., Gehin, C., Dittmar, A., and McAdams, E. (2012). EmoSense: An ambulatory device for the assessment of ANS activity-Application in the objective evaluation of stress with the blind. *IEEE Sensors Journal*, 12(3):543–551.

Matignon, R. (2007). *Data Mining Using SAS @ Enterprise Miner TM*. John Wiley and Sons.

Melillo, P., De Luca, N., Bracale, M., and Pecchia, L. (2013). Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. *IEEE Journal of Biomedical and Health Informatics*, 17(3):727–733.

Minsalud (2015). Informe Nacional de Calidad de la Atención en Salud 2015. page 217.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., and Marks, J. S. (2003). Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001. *JAMA*, 289(1):76.

Moncada, A. (2013). Toma de decisiones clínicas en atención primaria. *Rev Med Hered*, (24):319–323.

Mounce, S. R., Ellis, K., Edwards, J. M., Speight, V. L., Jakomis, N., and Boxall, J. B. (2017). Ensemble Decision Tree Models Using RUSBoost for Estimating Risk of Iron Failure in Drinking Water Distribution Systems. *Water Resources Management*, 31(5):1575–1589.

Mozumdar, A. and Liguori, G. (2011). Persistent Increase of Prevalence of Metabolic Syndrome Among U.S. Adults: NHANES III to NHANES 1999-2006. *Diabetes Care*, 34(1):216–219.

Murguía-Romero, M., Jiménez-Flores, R., Méndez-Cruz, A. R., and Villalobos-Molina, R. (2013). Predicting Metabolic Syndrome with Neural Networks. In Castro, F., Gelbukh, A., and González, M., editors, *Adv. Artif. Intell. Its Appl.*, pages 464–472, Berlin, Heidelberg. Springer Berlin Heidelberg.

Navarro, E., Lechuga, J., and Baquero, H. (2013). Barranquilla presenta alta prevalencia del síndrome metabólico. *Informativo UniNorte*, 11(84):4–5.

Navarro, E. and Vargas, R. (2008). Metabolic syndrome in the southeast of Barranquilla (Colombia). *Salud Uninorte,*, 24(1):40–52.

Navarro, E. and Vargas, R. (2012). Coronary risk according to Framinghan equation in adults with metabolic syndrome in the city of Soledad, Atlantico, 2010. *Revista Colombiana de Cardiología*, 19(3):109–118.

Navarro, E., Vargas, R., and Alcocer, A. (2016). Grasa corporal total como posible indicador de síndrome metabólico en adultos. *Revista Española de Nutrición Humana y Dietética*, 20(3):198.

Oh, B., Cho, B., Han, M. K., Choi, H., Lee, M. N., Kang, H.-C., Lee, C. H., Yun, H., and Kim, Y. (2015). The Effectiveness of Mobile Phone-Based Care for Weight Control in Metabolic Syndrome Patients: Randomized Controlled Trial. *JMIR mHealth and uHealth*, 3(3):e83.

ONS (2013). Enfermedad cardiovascular: principal causa de muerte en Colombia. (1).

Palacios, H. J. G., Toledo, R. A. J., Pantoja, G. A. H., and Navarro, Á. A. M. (2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. *Adv. Sci. Technol. Eng. Syst. J.*, 2(3):598–604.

Panchal, F. S. and Panchal, M. (2014). Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network. *International Journal of Computer Science and Mobile Computing*, 3(11):455–464.

Park, C.-K. and Kim, D. G. (2012). Historical Background. In *Progress in neurological surgery*, pages 1–12.

Perveen, S., Shahbaz, M., Keshavjee, K., and Guergachi, A. (2019). Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques. *IEEE Access*, 7:1365–1375.

Reaven, G. M. (1993). Role of Insulin Resistance in Human Disease (Syndrome X): An Expanded Definition. *Annual Review of Medicine*, 44(1):121–131.

Reyes, T., Xavier, A. F., and Naveiro, R. M. (2017). Systematic literature review of eco-innovation models : Opportunities and recommendations for future research. 149.

Rodríguez, A. S., Soidan, J. L. G., Gómez, M. J. A., Rodríguez, R. L., Alonso, A. d. Á., and Fernández, M. R. P. (2017). Metabolic syndrome and visceral fat in women with cardiovascular risk factor. *Nutr Hosp*, 34(4):863–868.

Rohanizadeh, S. S. and Moghadam, M. B. (2009). A Proposed Data Mining Methodology and its Application to Industrial Procedures. *Journal of Industrial Engineering*, 2(4):37–50.

Rückstieß, T., Osendorfer, C., and Van Der Smagt, P. (2011). Sequential Feature Selection for Classification. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. (2010). RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(1):185–197.

Shearer, C. (2000). J OURNAL Statement of Purpose E-Business and the New Demands on Data E-Commerce Places on Data Warehousing Technology WAREHOUSING. *Journal of Data Warehousing*, 5(4):13–22.

Smith, J. W., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 261–265.

Stephens, J. and Allen, J. (2013). Mobile Phone Interventions to Increase Physical Activity and Reduce Weight. *The Journal of Cardiovascular Nursing*, 28(4):320–329.

Stirrup, J. and Ramos, R. O. (2017). *Advanced Analytics with R and Tableau: Advanced Analytics Using Data Classification, Unsupervised Learning and Data Visualization*. Packt Publishing.

Stuckey, M., Fulkerson, R., Read, E., Russell-Minda, E., Munoz, C., Kleinstiver, P., and Petrella, R. (2011). Remote Monitoring Technologies for the Prevention of Metabolic Syndrome: The Diabetes and Technology for Increased Activity (DaTA) Study. *Journal of Diabetes Science and Technology*, 5(4):936–944.

Stuckey, M. I., Shapiro, S., Gill, D. P., and Petrella, R. J. (2013). A lifestyle intervention supported by mobile health technologies to improve the cardiometabolic risk profile of individuals at risk for cardiovascular disease and type 2 diabetes: study rationale and protocol. *BMC Public Health*, 13(1):1051.

Sun, Y., Wong, A. K., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719.

Tagle-Luzárraga, M., Gomez, F., and Tagle, A. (2007). Cardiovascular risk assessment using Framingham risk scoring in subjects with metabolic syndrome according to the ATP-III-NCEP, Framingham en sujetos con síndrome metabólico , definido por los crit. 54(4):211–215.

Tani, S., Iwata, M., Inada, H., Narazaki, H., Haraguchi, R., and Nakazawa, K. (2012). An online support system for promotion of a behavior change for persons with metabolic syndrome. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 1125–1129. IEEE.

United Nations, Department of Economic and Social Affairs, P. D. . (2015). World Population Prospects: The 2015 Revision, Key Findings and Advance Tables. Working Paper No. ESA/P/WP.241.

Urru, G. (2011). PRISMA declaration: A proposal to improve the publication of systematic reviews and meta-analyses. 135(11):507–511.

Vashist, S., Schneider, E., and Luong, J. (2014). Commercial Smartphone-Based Devices and Smart Applications for Personalized Healthcare Monitoring and Management. *Diagnostics*, 4(3):104–128.

Vashist, S. K. and Luong, J. H. (2019). *Point-of-Care Technologies Enabling Next-Generation Healthcare Monitoring and Management*. Springer International Publishing.

Vashist, S. K., Luppa, P. B., Yeo, L. Y., Ozcan, A., and Luong, J. H. (2015). Emerging Technologies for Next-Generation Point-of-Care Testing. *Trends Biotechnol.*, 33(11):692–705.

Vishal, K. and Kushwaha, A. S. (2018). Mobile Application Development Research Based on Xamarin Platform. In *2018 4th International Conference on Computing Sciences (ICCS)*, pages 115–118. IEEE.

Webster, J. and Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Q.*, 26(2):xiii – xxiii.

WHO (2013). *Global action plan for the prevention and control of noncommunicable diseases 2013-2020.*

Willers, J. and Hahn, A. (2013). Risk Assessment Using Two Different Diagnostic Tools: Metabolic Syndrome and Cardiovascular Risk Score (SCORE)—Data from a Weight Reduction Intervention Study. *Food and Nutrition Sciences*, 04(10):1028–1036.

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959):29–39.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

World Health Organization (2011). *Mhealth New horizons for health through mobile technologies*, volume 3.

Yan, A. F., Stevens, P., Wang, Y., Weinhardt, L., Holt, C. L., Connor, C. O., Feller, T., Xie, H., and Luelloff, S. (2015). mHealth Text Messaging for Physical Activity Promotion in College Students : A Formative Participatory Approach. 39(3):395–408.

Zimmet, P. Z. (1992). Kelly West Lecture 1991 Challenges in Diabetes Epidemiology–From West to the Rest. *Diabetes Care*, 15(2):232–252.