INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Finding patterns in cardiologic diseases using a data-driven approach

Filipa Isabel Ribeiro Gomes

Master's in integrated Decision Support Systems

Supervisor:
PhD, José Miguel de Oliveira Monteiro Sales Dias, Associated Professor with Habilitation,
ISCTE – University Institute of Lisbon

Cossupervisor:
MD, PhD, FESC, FAHA, Luís Afonso Brás Simões do Rosário, Assistant Professor,
Lisbon School of Medicine
University Cardiology Clinic

June, 2023

ISTA - School of Technology and Architecture

Finding patterns in cardiologic diseases using a data-driven approach

Filipa Isabel Ribeiro Gomes

Master's in integrated Decision Support Systems

Supervisor:
PhD, José Miguel de Oliveira Monteiro Sales Dias, Associated Professor with Habilitation,
ISCTE – University Institute of Lisbon

Cossupervisor:
MD, PhD, FESC, FAHA, Luís Afonso Brás Simões do Rosário, Assistant Professor,
Lisbon School of Medicine
University Cardiology Clinic

June, 2023

INSTITUTO UNIVERSITÁRIO DE LISBOA

Finding patterns in cardiologic diseases using a data-driven approach

Filipa Isabel Ribeiro Gomes

Acknowledgments

I would like to begin by expressing my sincere gratitude to all the people who contributed to the realization of this work, for without their support and guidance, this project would not have been possible.

First of all, I would like to thank the director of the Master's program, Professor João Carlos Ferreira, for the careful guidance, patience, and knowledge shared throughout this process. His dedication and guidance were fundamental for the development of this work and for my academic and personal growth. To my supervisor, Professor Doutor José Miguel Dias, for all the support and advice he gave me throughout this journey. To my co-supervisor, Professor Luís Rosário, who along this journey dedicated his time and shared his perspectives and experiences. The contributions of my advisor and co-advisor were crucial to the data analysis and to the validity of this study.

Secondly, I would like to thank the faculty members of the master's program, who provided an enriching learning environment. The invaluable insights and contributions during classes were an important piece to this work.

I want to thank my family and friends for all their emotional support and understanding throughout this academic challenge.

Finally, I would like to express my gratitude to all those who agreed to contribute in some way and took the time to participate in this project.

In summary, my research and writing of this master thesis has been enriched by the generosity, support, and contributions of many wonderful people. May this work contribute in some way to the academic community and to the advancement of knowledge in this area.

Resumo

A nível mundial, as doenças cardiovasculares (DC) são a principal causa de morte. Foram publicadas várias diretrizes para o tratamento das DC com o objetivo de melhorar a qualidade dos cuidados e reduzir os custos. Assim, é cada vez mais importante detetar e diagnosticar precocemente as doenças cardiovasculares.

Este estudo tem como objetivo construir um algoritmo que permita prever se o doente vai ultrapassar a sua frequência cardíaca. Para além disso, o objetivo foi construir um sistema de alerta que monitoriza o estado clínico do doente e, sempre que houver uma alteração, de acordo com alguns parâmetros, o médico recebe uma mensagem automaticamente. Este estudo teve como base um conjunto de dados do Hospital Santa Maria em Lisboa, obtidos através de Acordos de Prestação de Serviços Digitais desenvolvidos no âmbito do projeto FCT DSAIPA/AI/0122/2020 AIMHealth - Aplicações Móveis Baseadas em Inteligência Artificial para Resposta de Saúde Pública.

O método centrado nos dados seguiu a metodologia de Mineração de Dados (MD) CRISP-DM. Com base no conjunto de dados foi possível, seguindo esta metodologia, desenvolver um algoritmo de Aprendizagem Automática (AA) que pudesse prever antecipadamente se o doente iria exceder o intervalo interquartil da sua frequência cardíaca.

Verificámos que o nosso algoritmo de AA conseguiu prever problemas cardíacos em 90% dos casos e que o nosso sistema de alerta foi eficaz na deteção precoce de problemas cardíacos nos doentes. Este estudo demonstrou que a utilização de AA é uma ferramenta valiosa para detetar o agravamento do estado de saúde de um doente.

Palavras-chave: Análise de dados, Cardiologia, Saúde, Deteção Precoce, Sistema de Alertas, Inteligência Artificial.

Abstract

Globally, cardiovascular disease (CD) is the leading cause of death. Several guidelines for the treatment of CD have been published with the aim of improving the quality of care and reducing costs. Thus, it is increasingly important to detect and diagnose cardiovascular diseases early.

This study aims to build an algorithm to predict whether a patient will exceed their heart rate. In addition, the goal was to build an alert system that monitors the patient's clinical status and, whenever there is a change, according to some parameters, the doctor receives a message automatically. This study was based on a set of data from Santa Maria Hospital in Lisbon, obtained through Digital Services Agreements developed under the FCT project DSAIPA/AI/0122/2020 AIMHealth - Artificial Intelligence Based Mobile Applications for Public Health Response.

The data-centric method followed the CRISP-DM Data Mining (DM) methodology. Based on the dataset it was possible, following this methodology, to develop a Machine Learning (ML) algorithm that could predict in advance whether the patient would exceed the interquartile range of their heart rate.

We found that our ML algorithm was able to predict cardiac problems in 90% of the cases and that our alert system was effective in early detection of cardiac problems in patients. This study has shown that using ML is a valuable tool for detecting the worsening of a patient's health condition.

Keywords: Data Science, Cardiology, Health, Early Detection, Alert System, Artificial Intelligence.

Index

List of Figures

List of Tables

Glossary

AI- Artificial Intelligence

BPM- Beats per Minute

CD- Cardiovascular Disease

CRISP-DM Cross Industry Standard Process for Data Mining

DC- Doença Cardiovascular

DM- Data Mining

DNN- Deep Neural Network

KNN- K-Nearest Neighbors

LN- Logistic Regression

ML- Machine Learning

NCDS- Noncommunicable Diseases

PRISMA- Preferred Reporting Items for Systematic Reviews and Meta-Analysis

RF- Random Forest

SMTP- Simple Mail Transfer Protocol

SNS- Serviço Nacional de Saúde

SVC- Support Vector Classifier

TLS – Transport Layer Security

WoS - Web of Science

1. Introduction

Cardiovascular Diseases (CVD) in 2021 have been highlighted as the leading cause of death ata a global scale [1]. More than three-quarters of these fatalities occurred in developing countries [2]. This is because these countries have a shortage of health services and/or insufficient capacity, which causes late diagnosis [3]. It is estimated that the total number of deaths from these pathologies is likely to increase due to the growth and aging of the population [4]. Therefore, countries are adopting strategies to be able to address this problem [5]. A global action plan [6] between 2013 and 2030, referred to as Noncommunicable Diseases (NCDs), has been created that aims to encompass diseases such as cancer, diabetes, heart and lung disease and their risk factors. The main goal is to reduce mortality and increase collaboration and cooperation at national, regional, and global levels, enabling populations to achieve higher and better health outcome standards. The last statement [6] about this plan, in May 2022, noted that no country had reached all the goals that were set.

CVD have been also touted as the illnesses that are easiest to prevent and manage, while being the top cause of mortality worldwide during the previous ten years [4]. A quick identification of an illness is necessary for effective and thorough therapy [2]. There appears to be a pressing need for a reliable and organized method for detecting high-risk individuals and collecting data for prompt detection of CVD. Heart rate is an independent risk predictor for the onset of acute coronary even [7]. It has already been shown in several studies that increasing the heart rate by 10 beats per minute is associated with an increase in the risk of cardiac death by at least 20%, as can be seen in study [8].

Therefore, it is becoming increasingly important to face this growth using methods that allow for quick and efficient data collection and analysis [9]. In this context, the use of computational approaches such as Artificial Intelligence (AI), Machine Learning (ML) and Data Mining (DM), has been adopted and accepted in the scientific community. In fact, the use of automatic data collection systems in several industries enables daily gathering of massive volumes of data. The widespread use of information and communication technology, the development of cloud computing and the emergence of big data, led to the development of decision support systems and data mining technologies [7]. Increasingly we live in an age where data multiplies with each passing day. The amount of data generated in the world, by 2008, is the same amount of data that is now generated per minute [10]. IBM released a study that states

that every day 2.5 exabytes are created [11]. Making data mining and knowledge discovery systems relevant to a larger range of fields currently a trend [12].

Data mining techniques have been used by researchers to identify many non-communicable diseases ailments, such as diabetes [13], stroke [14], cancer [15], and heart disease [16]. Researchers have also attempted to use data mining methods to identify heart illness, considering the high prevalence of cardiovascular related mortality [17].

Researchers in data mining have long looked at how tools and technology might be used to perform data analysis in big data scenarios. Adopting data mining techniques in the medical area is crucial for accurately diagnosing and forecasting various disease conditions and comprehending healthcare data in depth. Among these applications, as shown in the article [18], we can highlight clinical data analysis, intended to enhance treatment practices and prevent hospital errors, early illness diagnosis, short term disease prognosis, disease prevention, and decrease in hospital mortality.

We can conclude that it is becoming increasingly relevant to identify heart disease problems as early as possible to improve patient outcomes. However, the symptoms of heart problems are painful, but they are often difficult to identify. This is because they are easily confused with sensations that people may experience after physical exertion or due to other situations.

Medical professionals face significant challenges in trying to identify these problems early. Due to the sheer number of patients, it becomes important that there are accurate and quantifiable reports to alert medical intervention. This would allow doctors to be called when needed and thus improve resource utilization. To address this challenge, we propose an accurate ML algorithm to predict a patient's risk of exceeding the normal heart rate and create an effective alert system to help medical professionals to act quickly. The data source for this study comprises a pseudo-anonymized multi-syndrome dataset of clinical data collected at Hospital Santa Maria, the largest Portuguese Public Hospital, located in Lisbon. Data was collected in the period of 01/01/2019 to 12/11/2021 that corresponds to the period of the Covid 19 outbreak, and the dataset is referred to as the HSM dataset. It includes data from 23122 patients and contains real-time clinical signals such as Temperature, Blood Oxygen Level (SpO2), Heart Rate. This dataset, whose schema includes 66 tables and occupies 75 Gbyte of data, was provided under the framework of the FCT project DSAIPA/AI/0122/2020 AIMHealth - Mobile Applications Based on Artificial Intelligence, aiming to contribute with a preventive approach for Public Health strategies in facing the COVID 19 pandemic situation, proposing a smartphone app and trustworthy Artificial Intelligence (AI) distributed and service-based platform, to identify symptomatic and asymptomatic patients as well as assess the risk of

exposure. The availability of the HSM dataset for research was approved by the Ethical Committee of the Faculty of Medicine of Lisbon, one of the project partners. The dataset is currently being accessed by ISCTE and Faculty of Medicine researchers, under a NDA and in the scope of the FCT AIMHealth project, which is also the case of this thesis.

The main contributions of our study are twofold: (1) the possibility of improving clinical decision making and the quality of care for patients with heart disease, with an intelligent computing approach based in ML and, (2) to combat the overload in hospital care by creating a system that generate alerts when there is an anomaly in the patient's health status, thus benefiting doctors and nurses.

1.1. Research Question

As mentioned before, heart problems are one of the leading causes of mortality worldwide. It is becoming increasingly important to be able to identify these problems early to improve patient outcomes. However, the symptoms of heart problems are subtle and can be difficult to detect. Medical professionals face significant challenges in trying to identify these problems early. In addition, with the high volume of patients that need to be monitored, it is easy to miss important information or not act quickly enough. This can result in treatment delays, prolonged hospitalizations, and less favorable patient outcomes.

To appropriately tackle this issue we can state our research question as follows: "how can we use ML algorithms to accurately predict the risk of a patient exceeding the normal heart rate and develop an alert system, based on pseudo-anonymized clinical data collected from patients at Hospital Santa Maria, so that medical professionals can act quickly and efficiently?".

To address this research question we followed a research methyodology, described in the next section. In short, we seek to develop an accurate ML algorithm, trained and tested on a pseudo-anonymized clinical dataset of Hospital Santa Maria patients, to predict a patient's risk of exceeding the normal heart rate and create an effective alert system to help medical professionals act quickly, that can be validated in the Hospital settings.

The main contribution of this study is the possibility of improving clinical decision making and the quality of care for patients with heart problems.

1.2. Research Methodology

For this dissertation, we started by following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) [19] methodology to perform a systematic literature review. This PRISMA method started with issuing a key research question "What is the state of the art on the use of Artificial Intelligence in aortic diseases data from Remote Monitoring Systems?". Fom this review we understood the major trends and gaps in the literature that could guide our research.

Increasingly we live in an age where data multiplies with each passing day. Thus, it is necessary to apply tools that help transform raw data (Big Data) and transform it into information and knowledge needed to manage organizations. An appropriate and popular data analytics methodology that emerged in 1996 to meet one the challenge of analysing Big Data, is the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology [20]. This is an effective and time-tested practice and is composed of six stages, that we adopted as our intelligent computing approach.

In the first stage, called Business Understanding, the goal is to understand the business, to understand the problem that needs to be solved, and to define the objective and the needs of the institution/organization. In our case this objective is to answer properly to our research question, whose context is the treatmets to patients with CDs at Hospital Santa Maria, during the COVID lockdown. In the second stage, the available data is collected, explored, preliminary analysed, and its quality is verified; this stage is called Data Understanding. In a third stage, Data Preparation, it is necessary to define data-centered modelling and analytics requirements, integrate data from different sources, and prepare data with enough quality for modelling and evaluation purposes. The fourth stage, Modelling, essentially involves selecting the most appropriate intelligent computing methods and algorithms that will properly answer the objective defined in the first stage (our resarchquestion). In a fifth phase called Evaluation, the model is applied and tested to verify if it meets the project's objective, using metrics adopted in the literature. Finally, Deployment, which as the name implies. is the computing implementation and delivery to the appropriate stakeholders of the results based on our predictive modeling and some heuristics, namely, the alert system that monitors the patient's clinical status, and whenever there is a change, according to some parameters, the doctor automatically receives a message, alerting for outliers or abnormalities predicted in the collected HR data.

For a better perception, Figure 1.1 represents the flow of the methodology that will be followed in this thesis.

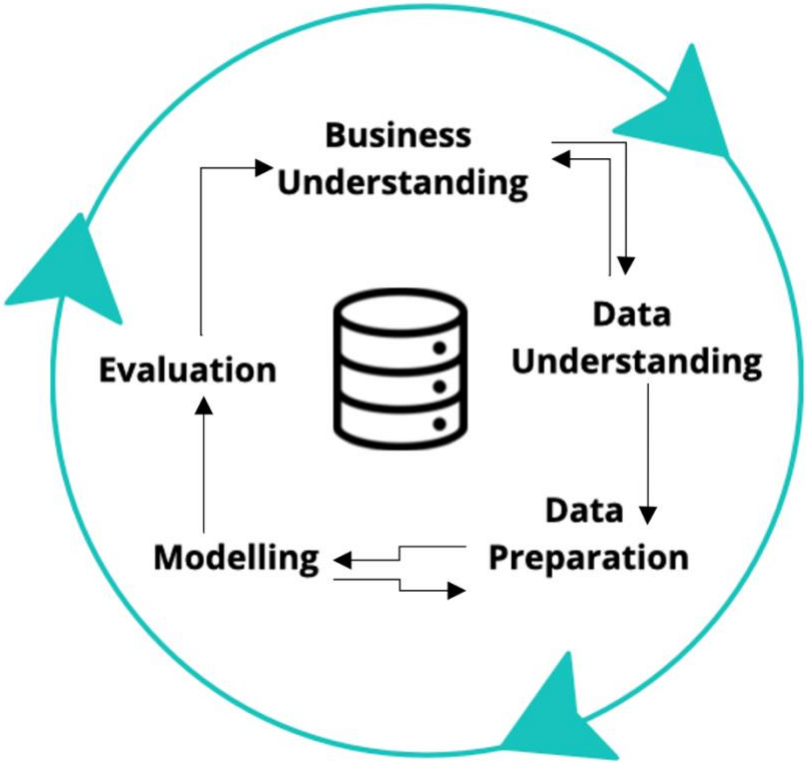

Figure 1.1 - Flow chart of the CRISP-DM

1.3. Outline of the Dissertation

This dissertation comprises four chapters (not including chapter one which is the introduction):

Chapter 2: The key to the whole process involves performing a systematic literature review, which was based on the PRISMA methodology, guided by the question: "What is the state of the art on the use of Artificial Intelligence in aortic diseases data from Remote Monitoring Systems?".

Chapter 3: This chapter's main objective was to analyze the obtained HSM dataset and to perform a data analysis using the CRISP-DM methodology.

Chapter 4: This chapter describes the design and development of an alert system. In this phase, a system was developed using Python that allows an alert to be generated automatically for the doctor, or nurse, that informs him/her that his patient's condition has worsened.

Chapter 5: Presents a discussion of the dissertation as well as projections for future work.

2. Literature Review

2.1 Search Strategy and Inclusion Criteria

This research was conducted on the 21st of May of 20223. Two repositories - Scopus and Web of Science (WoS) were used, and 45 articles were obtained. The criteria were only to choose articles published between 2020-2023 and the writing had to be in English. In addition, another limitation imposed was to choose only journal articles and articles and reviews.

2.2. Study Selection

The choice of articles was made by reading the title, the abstract and the keywords used by the authors.

2.3. Data Extraction and Synthesis

Zotero was used to store the articles, and articles were managed using Microsoft Excel and Microsoft Word, manging title, author, year, journal, subject area, keywords, and abstract.

To analyze and summarize the data, we performed a quantitative assessment. We searched the Scopus and WoS databases with the same keywords for research related to the concept of “Machine Learning”, "Artificial Intelligence" and “Data Analytics”, the target population of "Cardiovascular Diseases” and “COVID-19” with a context of the study of “Diagnosis”, “Medical data” and "Remote Monitoring" (see Table 2.1).

Concept	Population	Context	Limitations
Machine Learning	Cardiovascular Diseases	Diagnosis	Last 4 years
Artificial intelligence	COVID-19	Medical data	Only journal papers, articles and reviews
Data analytiscs		Remote Monitoring	
1 339 735 documents			
190 documents			
58 documents			
45 documents			

Table 2.1 - Definition of the Keywords Used to Search the Scopus and WoS Databases

Following PRISMA, we found, 45 documents. Next, it was necessary to perform a manual cleaning process. For this, duplicates were eliminated, same with documents whose title did not fit the search made and, finally, the complete article was analyzed and eliminated if it did not fit. In this way, it was possible to narrow the documents figure down to 33 documents. Our study took into special consideration the year, area, theme, and description of each paper.

Figure 2.1 shows the PRISMA methodology flow of the analyzed articles.

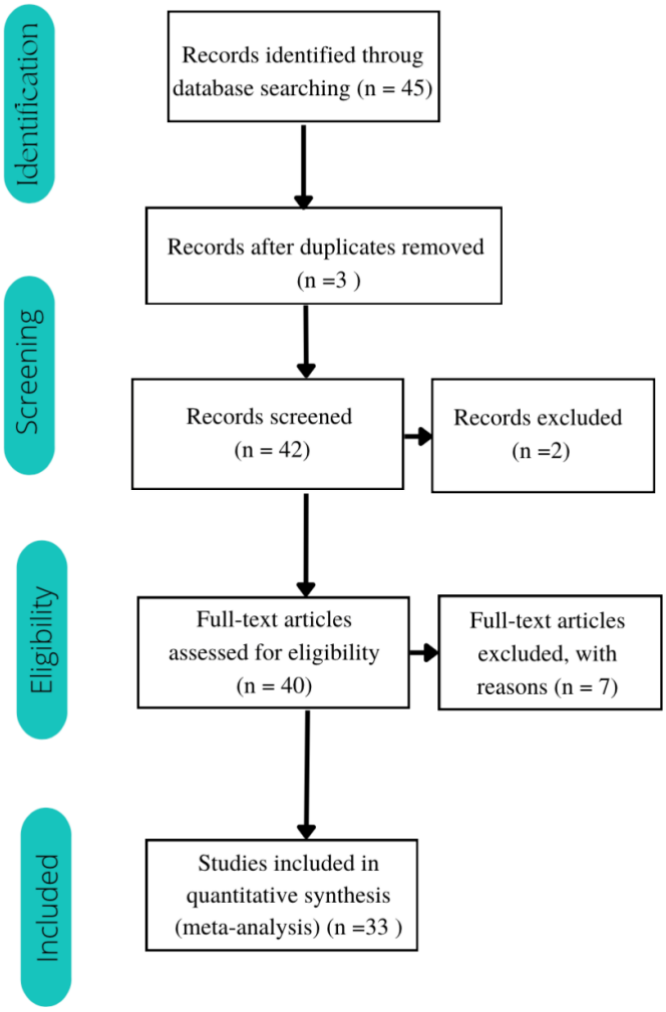

Figure 2.1 - PRISMA methodology flowchart

Figure 2.2 notes the clear upward trend in the number of papers, thus demonstrating importance the thesis topic.

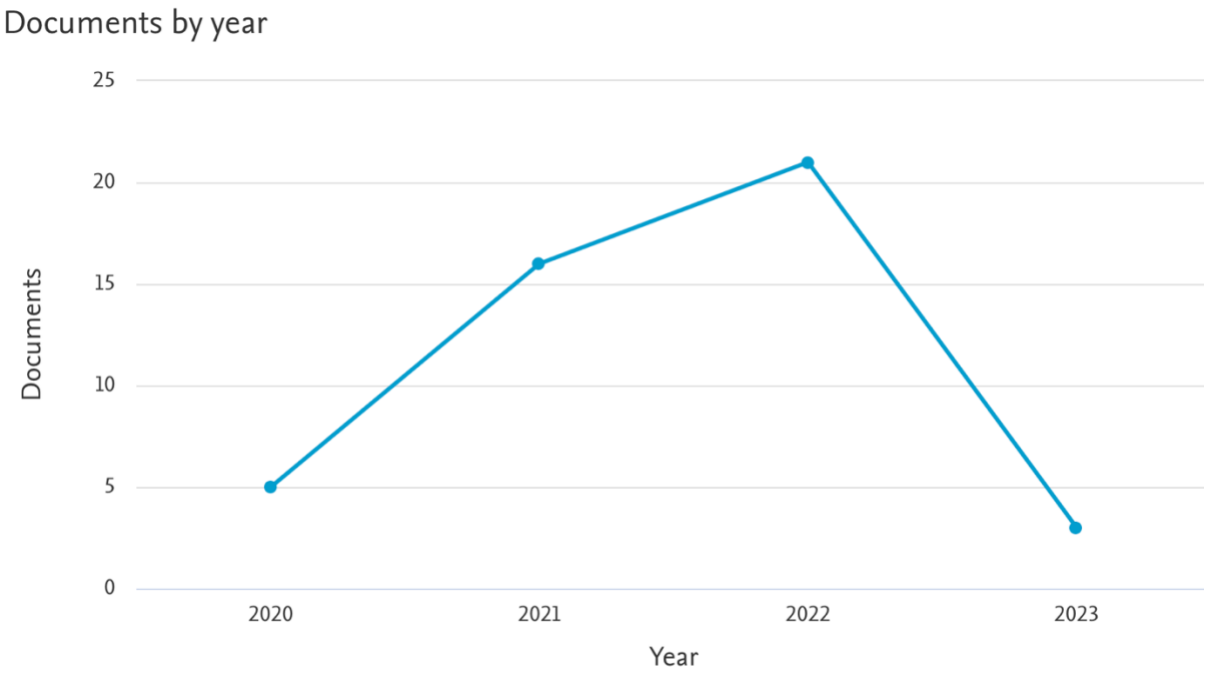

Figure 2.2 - Evolution by year of published articles

In a summarized way, in Table 2.2, it is possible, to analyze how the articles are distributed by topics.

Topic	Reference	# of Documents
Machine learning	[21]–[46]	26
Early Diagnostics	[21], [22], [25], [26], [28], [30]–[32], [37], [40], [44], [46]–[51]	17
Disease monitoring	[21], [22], [24], [25], [27], [29], [30], [48], [51], [73]	10
Remote Monitoring	[21], [22], [24], [25], [27], [29], [30], [48], [51], [73]	10
Disease relation	[21]–[23], [26], [27], [33], [48], [50]	8
Causes	[34]–[36], [38], [39], [46]	6
Evolution	[24], [28], [33], [36], [76]	5
DNN	[47]– [50]	4
Cluster analysis	[23], [76]	2
Review	[32], [43]	2
Post Illness	[27], [64]	2

Table 2.2 - Studies by Topics

Considering that one of the aims of this thesis is to identify which are the most used technologies in the analysis of aortic diseases, it is possible to observe in Figure 2.3 which are the main topics of this literature review.

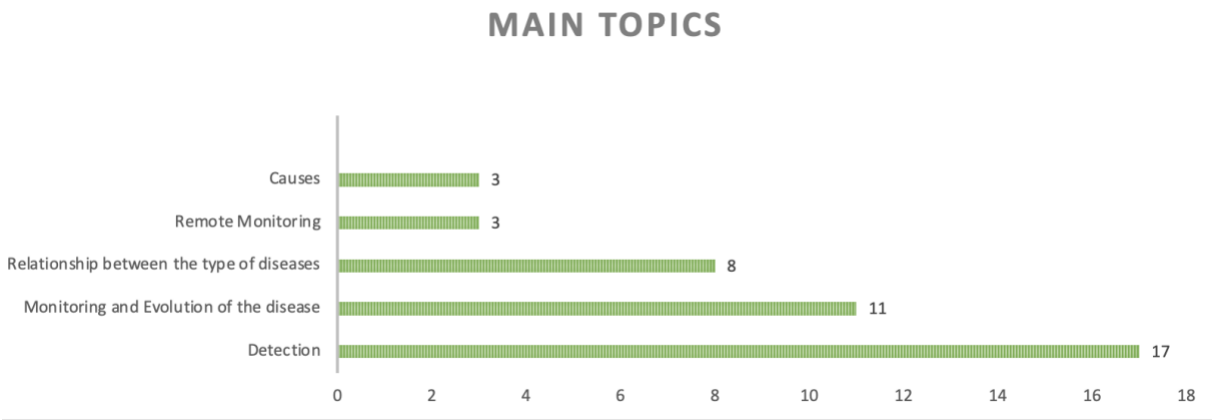

Figure 2.3 - Main Topics from the Literature Review

2.4. Literature Review Results

After analyzing the articles, it was possible to notice that there is a growth in studies using ML. In general, ML is a subset of AI, and Deep Neural Network (DNN) is a subset of ML. As mentioned in the article [25], there is a call from the scientific community for countries, especially those in development, to apply AI algorithms to curb the excess of deaths. In the literature review it is possible to see that, essentially, ML, DNN and Cluster Analysis (CL) are used.

Based on the defined keywords, we can conclude that the most used algorithm is ML representing 26 of the 33 selected papers. Studies that are based on the use of ML go on to demonstrate significant impact on the diagnosis and management of heart disease, including complications related to COVID-19, as well as in other areas of medicine. Studies [42], [47], [49], [51], [52] explore the use of ML techniques for early diagnosis of heart disease. The authors propose approaches based on signals from ballistocardiograph, respiratory effort, ECG, and photoplethysmography (PPG) signals to identify cardiac abnormalities and predict the risk of heart disease, including heart failure. In addition, some of these studies apply these techniques in the context of COVID-19, using ECG analysis to diagnose the presence of the disease.

Articles [46], [53], [60], meanwhile, address the broader application of ML in medicine. The authors explore how the combination of artificial intelligence and in vitro diagnostics can improve the accuracy and efficiency of medical diagnosis. In addition, some studies look at the use of artificial intelligence in precision medicine, which aims to provide personalized treatments based on genetic and individual patient characteristics. Also discussed are health

system planning and management during the COVID-19 pandemic, using artificial intelligence to support decision making and efficient management of health resources.

Papers, [47], [48], [54], [55], [66], explore the effects of COVID-19 on the cardiovascular system and propose Machine Learning-based approaches to diagnose and monitor cardiovascular complications in COVID-19 patients. The authors emphasize the importance of early diagnosis and ongoing management to improve clinical outcomes and health resource management during the pandemic.

Articles, [50], [63], [64] explore other applications of ML in healthcare, such as the use of mobile devices for cardiology care, the use of vocal biomarkers for medical diagnosis, and liquid biopsy technologies for hematological diseases. The authors highlight the potential of artificial intelligence in improving diagnostic accuracy, efficiency, and accessibility of healthcare in different areas of medicine.

The papers that focused on the use of DNN methods represent 4 out of 33 papers. Studies demonstrate the potential of AI and ML in diagnosing COVID-19 using ECG signals, also considering the presence of concomitant cardiovascular diseases. This study,[67], proposes an artificial intelligence-based approach to categorize ECG images of COVID-19, even in the presence of other cardiovascular diseases. The model uses DNN to identify specific patterns in the images that are indicative of the presence of COVID-19, considering the coexistence of cardiovascular conditions.

In paper [68], COV-ECGNET, a model based on deep convolutional neural networks for the detection of COVID-19 using ECG tracing images, is presented. The model is trained to identify distinct features in ECG tracings that are specific to COVID-19, considering the presence of concomitant cardiovascular diseases.

Study, [69], proposes ECG-iCOVIDNet, an interpretable artificial intelligence model to identify changes in the ECG signals of post-COVID individuals. The model uses DNN and interoperability techniques to identify patterns and abnormalities in ECG signals that are indicative of changes due to SARS-CoV-2 infection, also considering the context of cardiovascular diseases.

In paper, [70], a DNN model for diagnosing COVID-19 from images of paper ECG tracings is presented. The model can analyze the specific features in the ECG images and identify whether the individual is infected with COVID-19, considering possible cardiovascular comorbidities.

Finally, some authors use CA to study how the evolution of the disease in patients might be. Cluster analysis represents 2 of the 33 selected papers. Studies highlight the importance of

cluster analysis in identifying multimorbidity patterns and risk assessment in patients with COVID-19, particularly in the context of cardiovascular diseases. This study [44] uses cluster analysis to identify multimorbidity patterns and pre-existing conditions associated with in-hospital mortality in patients with COVID-19. Cluster analysis allows patients to be grouped based on similar characteristics, including cardiovascular disease, to identify groups at higher risk for mortality. The study seeks to understand how cardiovascular conditions and other comorbidities may influence clinical outcomes in patients with COVID-19.

Article, [76], discusses the use of cluster analysis and prognostic indices in the context of evaluating patients hospitalized with COVID-19 pneumonia. Cluster analysis allows for the identification of patterns of lung involvement and possible complications, including the impact on cardiovascular disease. In addition, the article discusses the use of prognostic indices based on CT scans to assist in assessing the risk of serious outcomes, also considering the presence of cardiovascular disease.

3. Data Mining

To address the first part of our how research question, namely, “how can we use ML algorithms to accurately predict the risk of a patient exceeding the normal heart rate (…), based on pseudo-anonymized clinical data from patients at Hospital Santa Maria, (…)?", we adopted the mentioned CRISP-DM (Cross Industry Standard Process for Data Mining) data mining methodology.

3.1. Business Understanding

In the CRISP-DM phase of Business Understanding, our goal is to understand the clinical context of production of the HSM dataset. Paper [77], help us unveiling such context, reflecting the reality of Portuguese hospitals. Study [78], shows that Portugal is the third OECD country with the most physicians. In the Portuguese public health system, doctors and nurses have a huge administrative burden, which is reflected in reduced available time to dedicate for patients. In contrast, the annual occupancy rate of hospitals in Portugal is around 84%, according to the SNS (National Health Service) report 2022 [79].

3.2. Data Understanding

In mid-May 2022 we started our access to the secure computing infrastructure where the raw HSM dataset was stored (the Iscte data center). We extracted a sample of the full dataset, with 600 Mbytes (66 csv files) of clinical information from 23122 patients. The main reason for choosing this sample is its computational efficiency. Exploratory data analysis showed that we were facing a fragmented data set.

Some of our sample dataset variables are gender, ethnicity, age, risk factors (pathologies, risks or allergies), the exact date of when the patient arrived at the hospital, the provenance (whether the patient arrived at the hospital alone, in an ambulance or through another hospital), the date of onset of symptoms, which type of professional they were seen by, the diagnosis given, the medications and treatments they will receive.

To better understand the data, we built a flow chart of the patient's journey through the hospital (depicted in Figure 3.1), by considering the analyzed tables and columns (variables).

The raw data accessed has many null values and no universal formatting in many variables, i.e., variables whose collection did not follow a strict formatting. There are a multitude of open fields which makes some of them difficult to analyze.

Figure 3.1 represents the hospital patient's journey. We derived such a journey by analyzing the provided data tables and their embedded links. Once the patient arrives at the hospital, he/she is entered at the "ADMISSION" table.

In the case of an emergency room or consultation, the department where the patient was admitted is registered ("DEPARTMENTS"), as well as the type of care ("ATTENDING TYPES"). Afterwards, the patient goes through a triage phase where the main complaints are registered in the "CHIEF COMPLAINTS" table. In this table we can find ID's for foreign tables such as: precautions to take ("PRECAUTIONS"); the type of allergies that the patient may have ("ALLERGIES"); the blood group ("BLOOD GROUP"); the patient's gender and ethnicity ("PATIENTS"). A first diagnosis of the patient is recorded in "ADMISSION DIAGNOSIS".

Occasionally, laboratory tests are performed, if applicable, and other necessary examinations are performed, whose information will be recorded entirely in the "LAB TESTS", "LAB RESULTS" and "ANALYSES" tables. The patient's diagnosis may also involve other types of tests that are registered in the "ORDERS" table. This table contains several corresponding foreign keys and describes: the duration and frequency of treatment ("SCHEDULES"); the type of route in which the patient received treatment ("ROUTES"); what unit of measure was used in the treatment ("UNITS"); the form of treatment ("FORMS"); types of treatment ("TREATMENTS"); and a scale of pain the patient is experiencing ("SCORE ITENS"). After the treatment is completed, the departments that accompanied the patient are recorded in the "ENVIROMENTS" and "ENCOUNTERS" tables, as well as any event that might have occurred. Finally, the patient may undergo a medical procedure that will be recorded in the "MEDICAL PROCEDURES TYPE" table.

3.3. Data Preparation

In the data preparation phase, an essential task is to determine the dataset to be used in our modelling approach, to achieve the main objective of this paper. In Table 3.1, you can analyze in detail which tables have been included for modelling and which will be excluded.

The inclusion of some variables was based on indications and insights provided by a medical doctor, specialized in cardiovascular diseases, in regular meetings that occurred since August of 2022. Our inclusion criteria are:

- Patient characteristics that may influence the development of cardiovascular disease. Tables such as: "ALLERGIES", "BLOODGROUPS", "COUNTRIES", "ETHNICITIES", "PATIENTALLERGY", "PATIENTPRECAUTION", "PRECAUTIONS" and "SEXES";
- Characteristics that relate to the diagnosis/treatment the patient has had. The case of tables like: "ADMISSIONDIAGNOSIS", "DIAGNOSES", "LABRESULTS", "LABTESTS", "MEDICALPROCEDURES", "MEDICALPROCEDURETYPES", "MEDICATIONS", "ORDERS" and " TREATMENTS";
- Characteristics that detail the patient's journey through the hospital. Such as the tables: "CHIEFCOMPLAINTS", "ENCOUNTERS", "EVENTDATA", "EVENTTYPES", "LOCATIONS" and "RTDATA";

On the other hand, the exclusion criteria were:

- Tables that were too “dirty”, containing many null values;
- Tables that did not follow a universal formatting, in other words, there had too many free text fields;
- Tables that did not add relevant information;
- Tables that were made up of values.

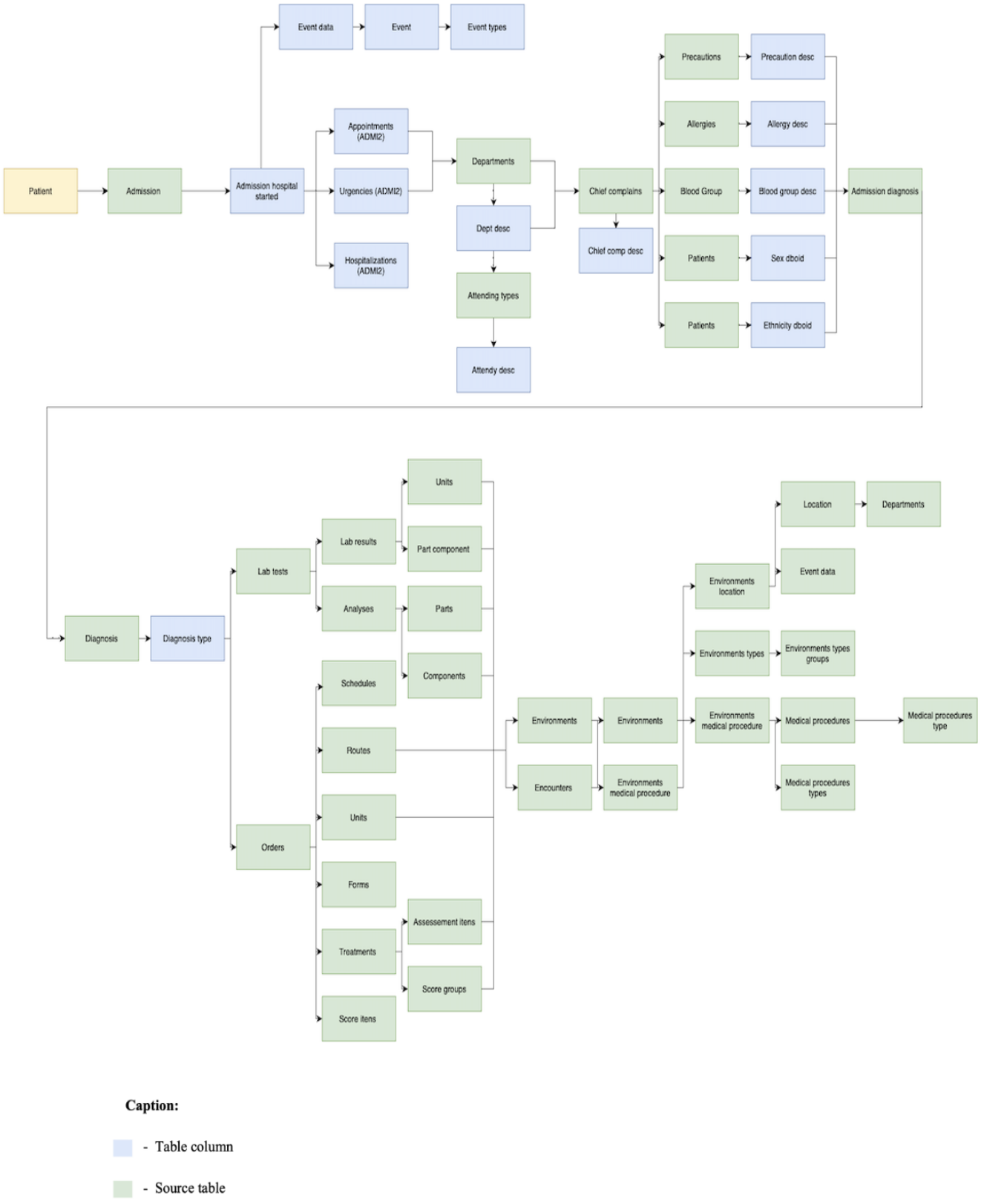

Figure 3.1 - Flow chart of the patient journey

In Table 3.1 we present a more detailed analysis of the tables that were included and excluded from our analysis.

Tables included in the analysis	Tables excluded in the analysis
ADMISSIONCHIEFCOMPLAINT	ADMISSIONAMBULATORYSTATUS
ADMISSIONDIAGNOSIS	ADMISSIONS
ALLERGIES	ALLERGYTYPES
ANALYSES	ASSESSMENTITEMS
BLOODGROUPS	ATTENDINGTYPES
CHIEFCOMPLAINTS	CATEGORIES
CNLCODES	COMPONENTS
COUNTRIES	DIAGNOSISTYPES
DEPARTMENTS	ENCNOTETYPENOTETYPEBLOCK
DIAGNOSES	ENCOUNTERNOTETYPE
ENCOUNTERS	ENVIRONMENTMEDICALPROCEDURE
ENVIRONMENTLOCATION	ENVIRONMENTS
ENVIRONMENTTYPEGROUPS	ENVIRONMENTTYPES
ETHNICITIES	EVENTS
EVENTDATA	FAMILIES
EVENTTYPES	FORMS
LABRESULTS	LABSOURCES
LABTESTS	NOTETYPEBLOCK
LOCATIONS	NOTETYPES
MEDICALPROCEDURES	ORDERTASKSTATUS
MEDICALPROCEDURETYPES	PARTCOMPONENT
MEDICATIONS	PARTS
ORDERS	PSICADATA
PATIENTALLERGY	PRECAUTIONTYPES
PATIENTPRECAUTION	ROUTES
PATIENTS	SCHEDULES
PRECAUTIONS	SCOREGROUPS
RTDATA	SCOREITEMS
SEXES	SELEC
TREATMENTS	STAFF
	STAFFATTENDINGTYPE
	STAFFTYPES
	TABLES
	TASKS
	UNITS
	SYSDIAGRAMS

Table 3.1 - Tables included and excluded from the analysis

Some tables contain important information and despite containing a substantial number of missing values and/or free text fields, were included in the analysis. Of the 24 tables that were selected for analysis, 14 contain null values.

The table "ADMISSIONS" is one of the most important for the analysis, although showing also missing values in some variables. It conveys relevant information about how the patient

was admitted to the hospital. In Figure 3.2 you can analyze the number of patients registered between the 1st of January of 2019 and the 12th of November of 2019, by department. Note that although it was stated earlier that the database contains information from 23122 patients the graph shows a higher count because a given patient may be admitted to the hospital more than once.

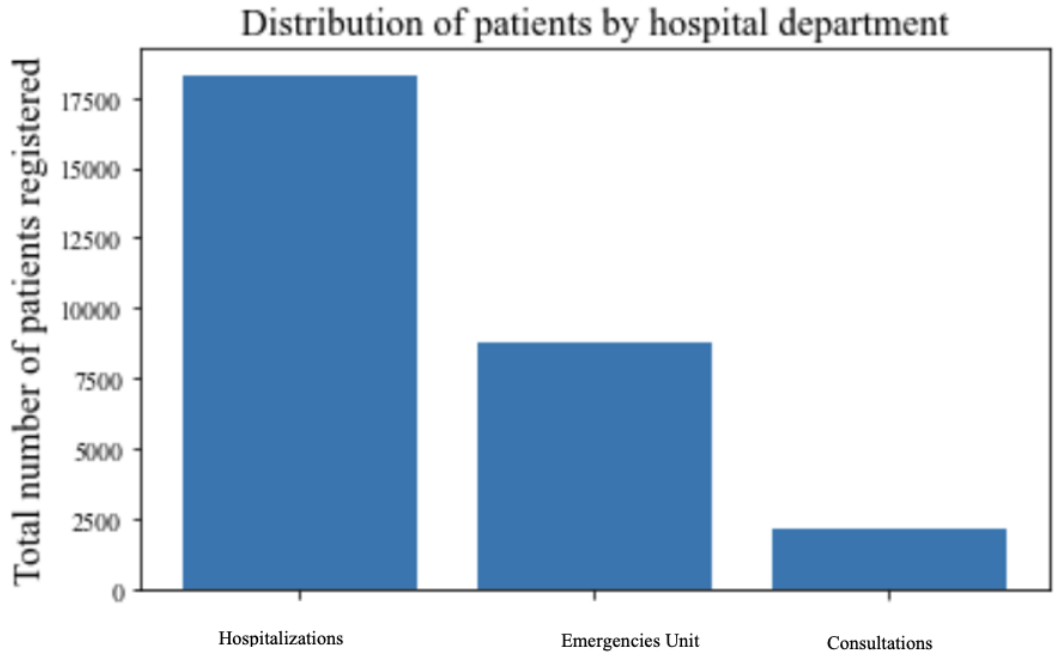

Figure 3.2 - Distribution of patients by hospital department

The aforementioned "ADMISSIONS" table contains some variables with null values, that convey information such as when the patient was admitted to the hospital, his weight, height, age and when he left the hospital. The variables for when the patient was admitted to the hospital is relevant, but we have in this table another variable (which does not contain nulls) for when the patient entered the system. Therefore, the variable “when the patient was admitted” can be eliminated. The variable for when he “left the hospital” corresponds to a table that we don't have access to, so this variable has all null values and had to be eliminated.

The tables "ANALYSES", "DEPARTMENTS", "ENCOUNTERS", "ENVIRONMENTLOCATION" and "LOCATIONS", contain respectively three, two, one, one and three variables with null values corresponding to foreign keys from tables that we do not have access to. Therefore, these variables were deleted.

The table "CNLCODES" contains coding information related to some other variables, such as “blood pressure”, “heartbeat”, “temperature”, and “pulse rate”. Therefore, this table

underwent a merge to associate the codes to the "RTDATA" table, despite including some null values.

The "DIAGNOSES" table allows us to know the diagnosis given to the patient. It consists of a code, its description and a foreign key from a table that was not used in the analysis. 1% of the data had null values and corresponding rows were deleted.

The "EVENTDATA" table has a high number of null values and non-conforming values, but it was chosen for the analysis because of the "STARTED" variable that is important to merge with other variables. Since this table was only used as a link to other tables, it remained as is.

The "EVENTTYPES" provides each event's category. This variable is relevant for us to understand the patient's table journey through the hospital.

The "MEDICATIONS" table deals with patients' medication and is composed of two variables with null values. These variables are not relevant because they only concern comments about the medications and their generic name. Therefore, these variables were not considered in our study.

The "PATIENTS" table deals with patient information and contains null values only for patient's birthday and ethnicity. The birthday variable contains 1.7% of null values, so these rows were deleted. The ethnicity variable shows 20% of null values, and the cor-responding rows were deleted.

The "RTDATA" table contains the most relevant information, namely the real-time clinical data, such as the patient's heartbeat, pulse rate, temperature, among others.

The "TREATMENTS" table only contains one variable with a foreign key with less than 1% of nulls. Thus, those rows with null values were eliminated.

After cleaning null values in the respective tables, we resort to building new tables. This step is important because it consists of integrating the data so that we have tables with only relevant information, for a better analysis. Although we will create tables that may have the same name, the goal is that these new tables have on the one hand, only the essential information from the original table, but also contain additional information from other tables that are relevant to the study.

We created the "Sitename_nhr" table, which corresponds to the description of the medical specialty where the patient is located. It corresponds to joining the tables "PA-TIENTS" ADMISSIONS", "DEPARTMENTS", and "LOCATIONS". It is composed of the “patient ID”, “admission ID,” “department type ID”, “department description”, “location”, and

“hospital code” variables. The purpose of this table was to help us understand the patient hospital journey since admission.

Next, we created the "Country" table, with “patient's id”, “country id”, and “description” variables. It is possible to see, in Table 3.2, that more than 98.5% of data correspond to unknown values. Considering such high value, corresponding rows were not deleted. The purpose of creating this table was to later be linked to a mother table (also referred to as the Main table).

Value	Frequency (%)
Unknown	98.5%
Portugal	1.4%
Brazil	<0.1%
United Kingdom	<0.1%
Germany	<0.1%
Cape Verde	<0.1%
Denmark	<0.1%
France	<0.1%
Senegal	<0.1%

Table 3.2 - Distribution of patients by country

The "Gender" table was also built, composed of the “patient's id” and “gender”. In Table 3.3, we notice that the male gender is more predominant.

Value	Frequency (%)
Male	52.7%
Female	45.6%
Unknown	1.7%
Undetermined	<0.1%

Table 3.3 - Distribution of patients by gender

An "Admission_sympt_date” table was included, to enable the analysis of symptoms that the patient had at the time of admission, namely, understanding the date the symptom occurred, the patient's precautions, and the environment the patient was in.

We also wanted to understand the date of the first admission in any facility, thus leading us to the creation of "Admission_any_date" table. Looking at Table 3.4. it is possible to conclude that more than half of the admitted patients went to the operating room.

Value	Frequency (%)
Operating room	59.0%
UCPA	10.6%
Neutral hospitalization	8.1%
Cardiothoracic ICU – Intensive Care Unit	5.3%
Intensive Care Medicine Service	3.9%
Neuro ICU	3.8%
Pediatric ICU	2.9%
UCIR- ICU	2.0%
NICU	1.4%
UTIC-ICU	1.4%
Other values (4)	1.5%

Table 3.4 - Distribution of patients by service

Next, we created the "Caut_patients" table which consists of a merge of the “pre-caution type ID”, its “description”, the “precaution ID” and its “description”, the “patient's precaution ID”, the “start of the precaution”, when the precaution ended, the “patient ID”, “comments”, the “patient's birthday”, “blood group”, “ethnicity”, and “gender”. Since this was a complete merge without restrictions, some repeated variables had to be deleted. Variables that did not add relevant information, such as “birthday”, were also deleted. The “started” and “ended” variables were also eliminated because they had too many null values.

In addition, we built the "Chief_comp" table. This table is composed of the patient ID, the admission ID, and the chief complaint. A "Precaution" table was developed, which is composed of the “patient ID”, the “precaution”, and the “type of precaution” variables. Most of the data values are unknown, as can be seen in Table 3.5.

Value	Frequency (%)
Unknown	95.2%
Background	3.2%
Anesthetic complications	1.1%
Other	0.2%
Biological/Infectious Risk	0.1%
Bone or/and joint changes	0.1%
Spinal or epidural puncture	0.1%
Venipuncture	<0.1%

Table 3.5 - Distribution of patient’s precautions data

We created also a "Diagnosis" table, composed of variables like “patient ID”, “admission ID” when the admission took place, “diagnosis”, “diagnosis code”, and “type of diagnosis”. 64.9% of the diagnoses are unknown, as can be seen in Table 3.6.

Value	Frequency (%)
Unknown	64.9%
Acute Appendicitis	0.6%
Dyslipidemia	0.5%
HTA	0.4%
Phimosa	0.4%
Atherosclerosis	0.4%
Neoplasia Bladder	0.4%
Hydronephrosis	0.3%
Respiratory Insufficiency	0.3%
Aortic Valve Diseases	0.3%
Other Values (2429)	31.6%

Table 3.6 - Distribution of patients by type of diagnosis

Then the "Procedures" table was structured. It includes “patient ID”, “admission ID,” “beginning” and “end” of the admission, “episode ID”, when the episode “started” and when it “ended” (the end of the episode always corresponds, for this data, to the patient's discharge).

Next, the "Events" table was designed, which is composed of the “patient ID”, “admission ID”, the “event start”, the “event type”, and the “event description”.

Finally, the last table was created, which is characterized as "Main table", corresponding to a mother table. This table includes variables that are the most relevant for cardiovascular disease analysis, according to a medical doctor specialist in cardiovascular diseases and one of the authors of this paper, namely, “patient's ID”, “ethnicity”, “blood group”, “gender”, “event start”, “temperature”, “systolic” and “diastolic” pressure, and real-time data variables such as “oxygen saturation”, “oxygen saturation”, “respiratory rate”, “heart rate”, patient's “pulse rate”. Since values were missing in some of these variables (“temperature”, “systolic” and “diastolic” pressure, “oxygen saturation”), those were eliminated.

3.4. Modeling

In the fourth phase of the CRISP-DM methodology, named Modeling, we created our predictive model. For the basis of the model, we used data from the "Main_table". As mentioned in subsection 3.3. Data preparation, this table contains information about each patient in terms of real time registry, respiratory rate, heart rate, pulse rate, ethnicity, blood group, and gender. It was necessary to convert the “respiratory rate”, “heart rate”, and “pulse rate” variables to numeric values, while the “ethnicity”, “blood group”, and “gender” variables were ignored in the model. As mentioned earlier, the goal was to build an algorithm that could predict whether the patient would exceed the interquartile range of the heart rate, based on the values of the

input characteristics. To this end, we created a Boolean variable, "Binary_ FC", which had a value of 1 if it was within the interquartile range and a value of 0 if it was outside of that range.

We performed an exploratory data analysis to understand what the ideal time interval would be and if it made sense to use the complete database, with all the patients, or perform the prediction model for a specific patient. The "Main_table" contains (real-time) time-series variables, including the “heart rate”, that provide instantaneous clinical data measurement. However, such variables contain records with varying sampling rates. For example, there are sets of records whose sampling rates are in the minutes range, while others in the second’s range. To get around this problem we needed to standardize the sampling rate of the time-series variables of interest. Note that we always performed the average of the records in the time interval associated to selected sampling rate. Several tests were performed using the Python statsmodels library [80], aiming to understand what would be the time interval that could provide us with a better correlation between the variables.

Figure 3.3 and Figure 3.4, depict the autocorrelation at sampling rates of, respectively, 30 and 15 minutes for the "Binary_ FC" variable of interest, and based on all patient’s data.

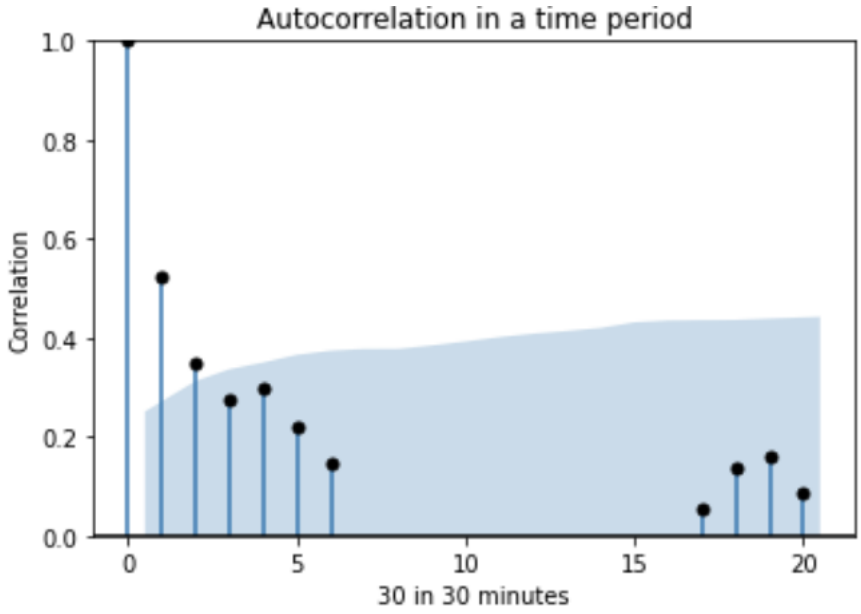

Figure 3.3 - Autocorrelation in a 30-minute interval

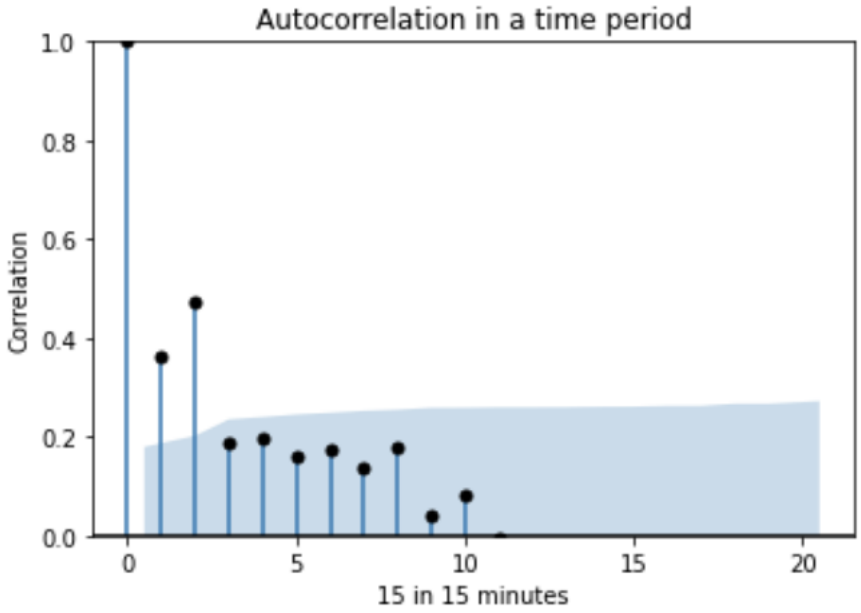

Figure 3.4 - Autocorrelation in a 15-minute interval

In Figure 3.5 and Figure 3.6, we visualize the autocorrelation at sampling rates of, respectively, 30 and 15 minutes for the "Binary_ FC" variable of interest, for a specific patient that was selected as the one with the highest values.

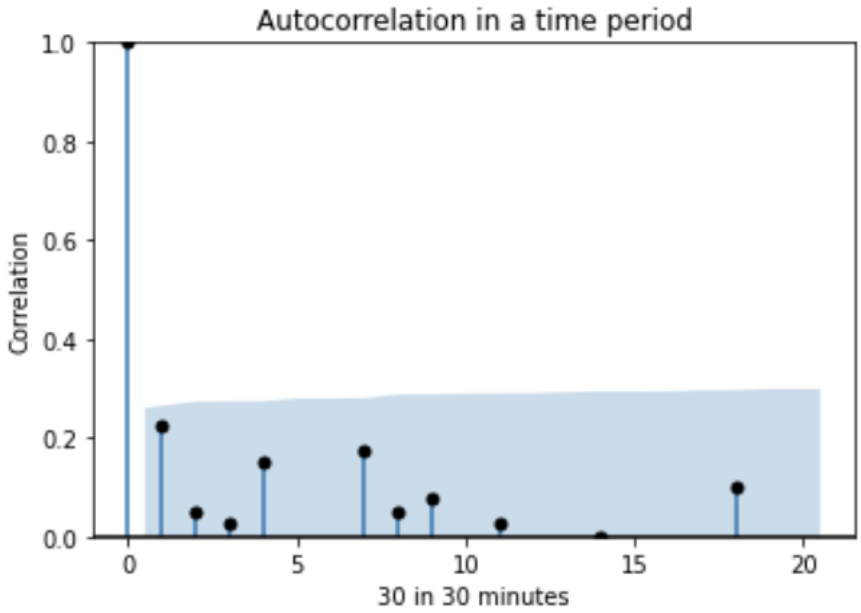

Figure 3.5 - Autocorrelation in a 30-minute interval

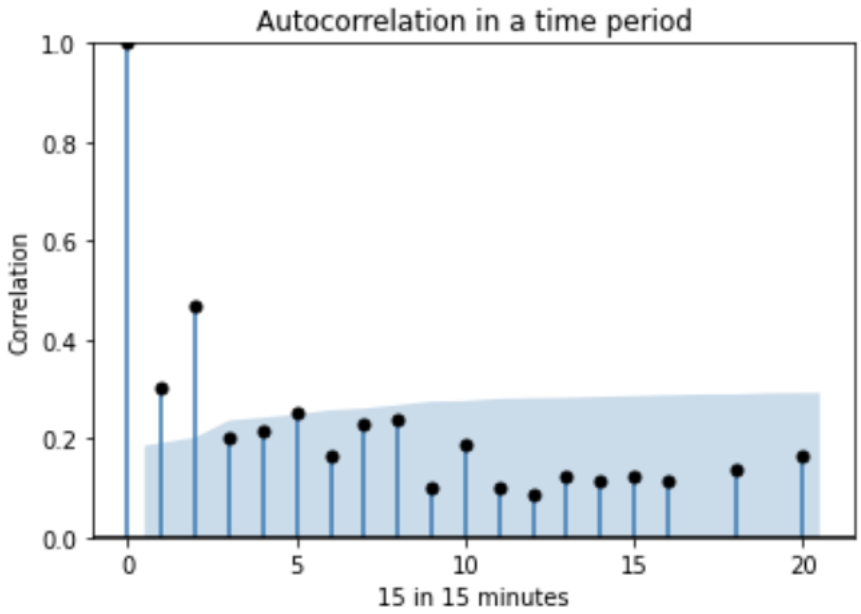

Figure 3.6 - Autocorrelation in a 15-minute interval

By looking at the previous graphs, it is possible to get an idea of what would be a good choice of the data sampling rate for our variable of interest, for modeling purposes. We selected a specific patient with data sampled every 15 minutes. The choice was made bearing in mind that the patients would have different inter quartile amplitudes, so the model should ideally focus on a specific patient. Regarding the time interval, we chose sampling intervals of 15 minutes, because it was the one that presented the highest correlation. After this exploratory analysis, we concluded that our solution, able to properly address our research question, should select an algorithm that would predict a categorical Boolean variable ("Binary_ FC "), our target variable, informing the analyst if the patient would exceed or not the interquartile range of heart rate, 1 hour and 15 minutes in advance. Each X stands for a 15-minute interval, so we equaled the X to 5 to set the prediction interval to 1 hour and 15 minutes.

The "train_test_split" library Python from scikit-learn [81] is used to split the data into training and test sets, for machine learning modelling purposes. The library randomly splits the data into two subsets based on the "test_size" parameter, which in this case is set to 0.3. This means that 30% of the data will be used for testing, and the remaining 70% will be used for training.

The selection of which algorithm to use to build the model was one of the most important decisions in the process. Our choice focused on supervised ML type of models, since we would be dealing with predictions based on collected data and insufficient data. ML models such as

Naive Bayes (NB) [82], Logistic Regression (LR)[83], K-Nearest Neighbors (KNN) [84], and Random Forest (RF) [85] were applied.

For the applicable performance evaluation, we considered the standard accuracy, recall, F1 score and precision metrics, typically used for categorical variables prediction. In medical terms it is more relevant to detect false negatives with high accuracy, because if we diagnose that a patient doesn't have an infectious disease when in fact he or she does, we will have a consequence of more infected people. Thus, the most important metric to consider in this analysis is recall, which gives us out of the total positive, what percentage are predicted positive, since we want to minimize false negatives.

For the NB algorithm there is only one hyperparameter to tune which is the "*var_smoothing*" that consists of a smoothing for the variance estimation. In Figure 3.7, the values of the "*var_smoothing*" parameter are constant providing a maximum recall of 0.8333.

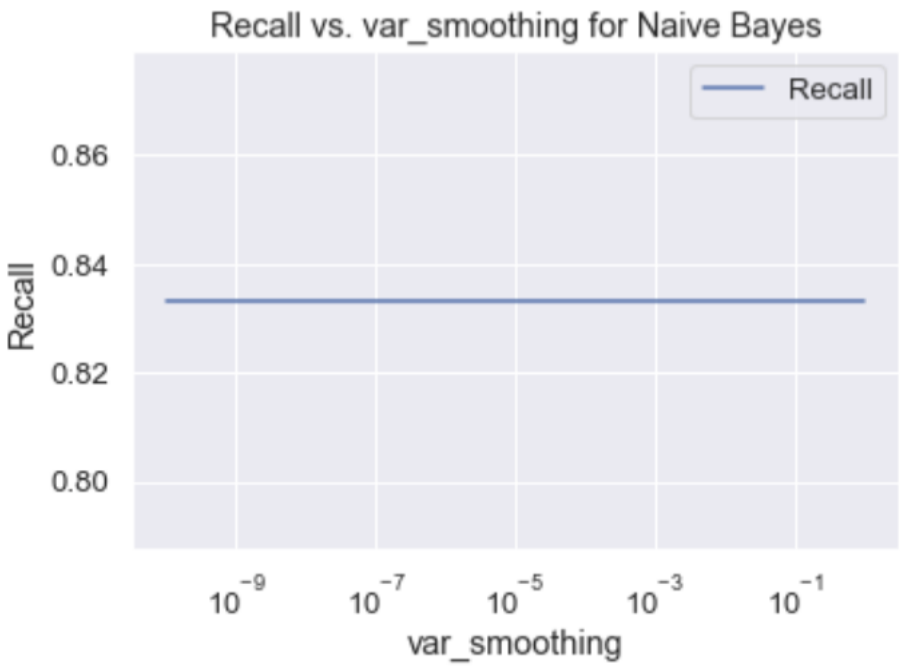

Figure 3.7 - Determining the hyperparameter for the NB algorithm

For the LR algorithm, it was necessary to define the maximum number of iterations that could transmit the best recall value. In Figure 3.8, you can see that starting at 50 iterations we reached the maximum recall.

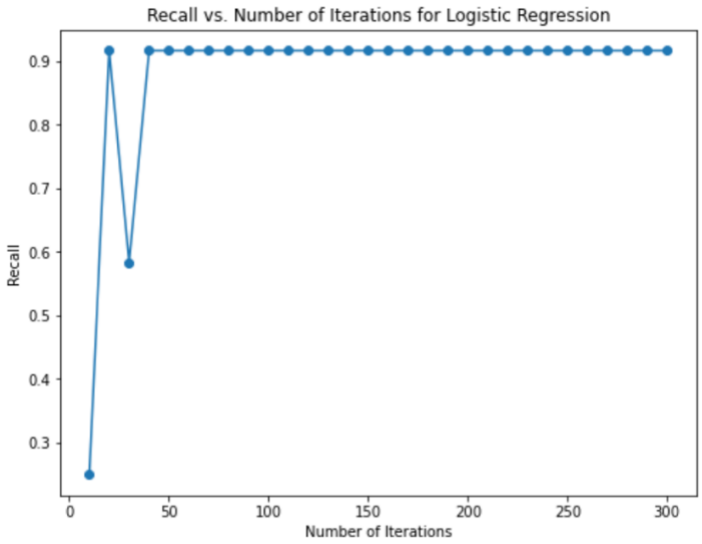

Figure 3.8 - Determining the maximum number of iterations that gives the best recall for the LR algorithm

For the KNN algorithm, we needed to define the optimal K. When fitting a KNN model to a data set, the algorithm stores all training data and corresponding target values in memory. To make a prediction for a new data point, the algorithm finds the K training data points that are closest to the new data point in terms of Euclidean distance, and then makes a prediction base on the most common target value among the K closest neighbors. In Figure 3.9 you can see that the optimal number of K is equal to 27.

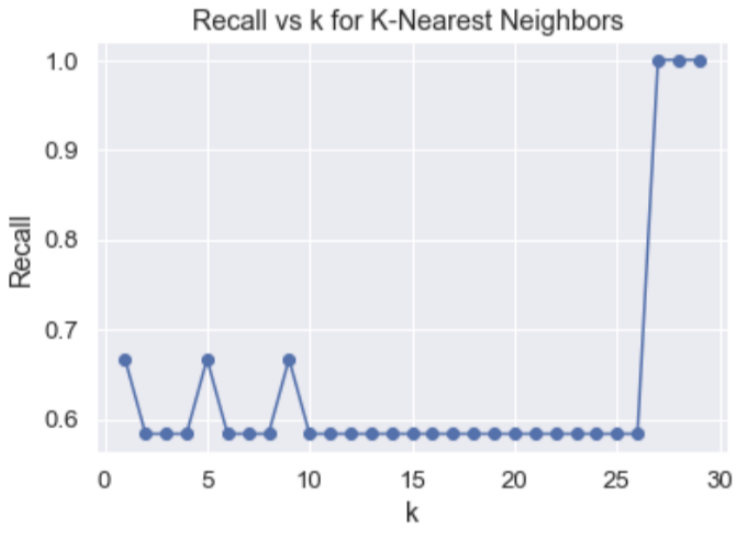

Figure 3.9 - Determining the best k for the KNN algorithm

Finally, in the RF algorithm it is important to find the hyperparameters that can maximize recall. The hyperparameters to establish for this algorithm are the number of estimators and the random state. The number of estimators is the hyperparameter that constructs the number of decision trees. A larger number of decision trees can improve the model but increases the computational cost and the overfitting echo. In turn, the random state allows us to determine the initialization of decision trees and their sampling of features and data points. As can be seen in Figure 3.10, to obtain the maximum recall the number of estimators was set to 50 and the random state was set to 30.

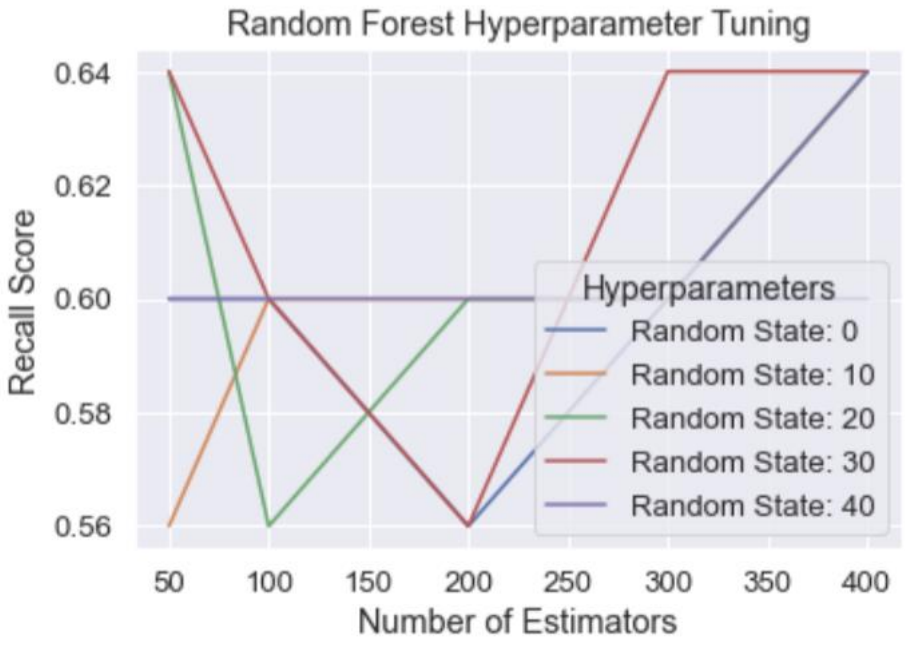

Figure 3.10 - Determination of the best hyperparameters that can maximize recall for RF

3.5. Evaluation

The last phase of the CRISP-DM methodology is to understand if the modeling results that were obtained are considered relevant, considering our research question and the selected solution, namely, building a machine learning algorithm able to predict if a patient would exceed the interquartile range of their heartbeat.

To evaluate the model applied to our case of Boolean variable prediction, several evaluation metrics need to be considered. In our model, the algorithm takes as input numerical variables such as “pulse rate” and “blood oxygen”, as well as Boolean variables such as “gender”,

“ethnicity”, and “blood group”. As mentioned, the goal is to minimize false negatives, making recall the most important metric to consider.

Table 3.7 depicts the results achieved with different evaluation metrics, over all the mentioned algorithms. The ratio of records, as mentioned above, is 70% to 30% which corresponds to 51 and 23 records, respectively.

	NB	LR	KNN	RF
Accuracy	60,87%	73,91%	52,17%	56,52%
Precision	58,82%	68,75%	52,17%	57,14%
Recall	83,33%	91,67%	100%	66,66%
F1 score	68,97%	78,57%	68,57%	61,54%

Table 3.7 - Performance of different algorithms with different metrics

When evaluating the task of predicting if a patient will exceed the interquartile range of their heartbeat, Logistic Regression was chosen as the best algorithm, by analyzing table 8. However, it is important to understand why the other three algorithms were excluded. NB shows an accuracy of 60.87%, precision of 58.82%, recall of 83.33%, and F1-score of 68.97%. While its recall was higher than LR, its precision was much lower, making it less suitable for this medical application. NB also assumes independence between features, which may not hold and could lead to inaccurate predictions.

KNN achieved a recall of 100%, but an accuracy and precision of only 52.17%, and an F1-score of 68.57%. KNN's perfect recall can be attributed to its ability to identify the nearest neighbors to a given observation, but its low precision suggests that it may classify many observations incorrectly. KNN also suffers from the curse of dimensionality, where the performance deteriorates as the number of features increases.

RF had an accuracy of 56.52%, precision of 57.14%, recall of 66.66%, and F1-score of 61.54%. RF is known for its interpretability and ability to handle both continuous and categorical (in our case) variables. However, in this case, it performed worse than LR in terms of recall, precision, and F1-score.

In contrast, LR had an accuracy of 73.91%, precision of 68.75%, recall of 91.67%, and F1-score of 78.57%. Furthermore, it is an interpretable model, which is a desirable feature for medical applications where understanding the decision-making process is important. LR can

also handle both continuous and categorical variables and is less prone to over-fitting than other complex models.

In conclusion, LR was the best algorithm for predicting if a patient will exceed the interquartile range of their heartbeat based on its high recall, interpretability, and ability to handle both continuous and categorical variables. NB, KNN, and RF were excluded due to their lower performance in terms of recall and/or precision, as well as their limitations in handling the dataset's features.

4. Design and Development of The System

To address the second part of our how research question, namely, “how can we (…) develop an alert system, based on pseudo-anonymized clinical data from patients at Hospital Santa Maria, so that medical professionals can act quickly and efficiently?”, we developed a prototype alert system to mitigate the workload of healthcare professionals, particulary doctors and nursing staff. The system is programmed to alert the doctor, or nurse, in charge of a given patient, via email, as soon as his/her heart rate and/or pulse values, measured in real time, deviates from what is acceptable. The system is based on the previously developed ML algorithm and metaheuristics. The goal was to develop a system that is versatile and dynamic, and therefore capable of being used anywhere in the world. Because of its versatility, the system also supports data from the ML algorithm making it possible to send alerts based on predictions that the algorithm makes.

For the metaheuristics part of the system the variables heart rate, pulse rate, temperature, systolic pressure, and diastolic pressure. The records in the table are from 1226.

During the design of this system, the Pandas Data Frames [86], NumPy [87] and SMTPLib [88] python packages were used. The first two allowed for data processing so that it was possible to use only the data necessary for the intended purpose, while the last one allowed the goal of automating the sending of notifications to health professionals.

4.1. Process

We used the SMTP (Simple Mail Transfer Protocol) package, due to its ease of implementation and versatility, as well as not being demanding in terms of resources. Then the necessary configurations were made, including determining from which email address the alert would be sent. For system development purposes, an email address was created specifically for that purpose (alertshsm@gmail.com), simulating a real-world environment, where a health professional would receive an email from the hospital. The address was created in the Gmail email platform.

Then, from the set of patient data, the system filters the information and will select the five relevant variables from the table, using Pandas. The selected variables are: "Heart_Rate", "RT_O2", “RT_Temp”, “RT_Systolic” and “RT_Diastolic”. Afterwards, each of the variables is converted into arrays using NumPy so that we can more easily treat the data and make it

uniform for further analysis. The values were also converted to integers, because it was considered that this would be the most appropriate format considering the information they refer to.

The next step was to train the algorithm previously created in chapter 3. Whenever an abnormal event is predicted to happen the algorithm will issue an alert to the physician that the patient may exceed his/her heart rate. Nevertheless, where we don't have access to annotated data the system will average each variable, for the patient in question, to be used as a reference value for later comparison with the real-time values. For now, the data from the HSM dataset does not contain information on the systolic and diastolic blood pressures and the respective temperature of the patients and therefore these variables were not included in the analysis, although the system was designed to support them.

To fix an acceptable interval that the values in each variable could assume, before issuing an alarming situation, a medical doctor specialist in cardio-vascular diseases and one of the supervisors of this thesis (Prof. Dr. Luís Rosário), decided that it would be optimal to fix the interval at 10% variation. In other words, the values were considered abnormal if they were between 10% below or above the average.

Finally, we created five loops to analyze each array and automate the process. These were created so that the system checks if the value registered in each of the five variables selected is within the acceptable range. However, if at any time the registered value is outside that range, the system sends an immediate alert to the email address determined for that purpose. To protect against spam and unnecessary warnings, a cooldown timer was set at 10 minutes from the time that an email was sent. This allows for a doctor or a nurse to have time to check the patient and act without having to receive multiple warnings concerning the same crisis.

Given that there are five variables to monitor it is imperative that the system can perform the checks on each column at the same time. To allow this, we imported and used the Threading python [89] module, making the execution of multiple loops simultaneously possible.

When developing this alert system, importance was also given to the protection of the personal data of the parties involved, such as email addresses, names, clinical data and even information related to the hospital. To maintain the confidentiality of this data, an increasingly important concern nowadays due to the rise of cybercrime, we used the TLS (Transport Layer Security) protocol [90], to encrypt the data and communications used. This protocol is the one recommended when it comes to data encryption in communications, at the present time. Preceded by the now obsolete SSL (Secure Sockets Layer) [91], it uses symmetric and asymmetric encryption to prevent information from being compromised by a third party that

should not have access to it. However, it is important to note that this protocol protects the transmission of data, preventing access to it during the exchange between systems as well as its adulteration. It does not protect unauthorized data access when data is already inside the computer systems, which means that the data does not remain encrypted outside the transmission channel itself. The TLS security protocol uses symmetric cryptography, in which the two parties involved in the trans-mission share a common key used for encryption and subsequent decryption of the in-formation; and asymmetric cryptography, where there is a public key used for encryption, but in decryption, a private key mathematically derived from the first (the public key), but specific and unique to each user, is used. The combination of these two methods allows this to be a secure tool in data transmission.

4.2. Demonstration of the Process

An important step in the work involves testing and evaluating the entire process that has been carried out, to see if the goal has been successfully achieved, that is, to properly answer to our research question.

For data protection purposes, fictitious names were used in the system demonstration. Thus, we used the pseudo-anonymized ID "1742575894659769552" whose fictitious name is Rita.

Rita was admitted to the hospital on January 7th, 2019, at 9:29pm, but the first monitoring was performed at 11:35 pm of the same day. The patient has 1072 records, where the observed interval between measurements was between 3 and 7 minutes. Figure 4.1 shows, hour by hour, the evolution of the patient's heartbeat and pulse rate. Since several values had been recorded in one hour, we decided to compute an average of the heart rate and pulse rate for that hourly period. These two variables had 20% null values, but these were replaced by the average of the respective variables. Rita's average calculation of the variables was 74.4 bpm and 73.7 bpm for heart rate and pulse rate, respectively.

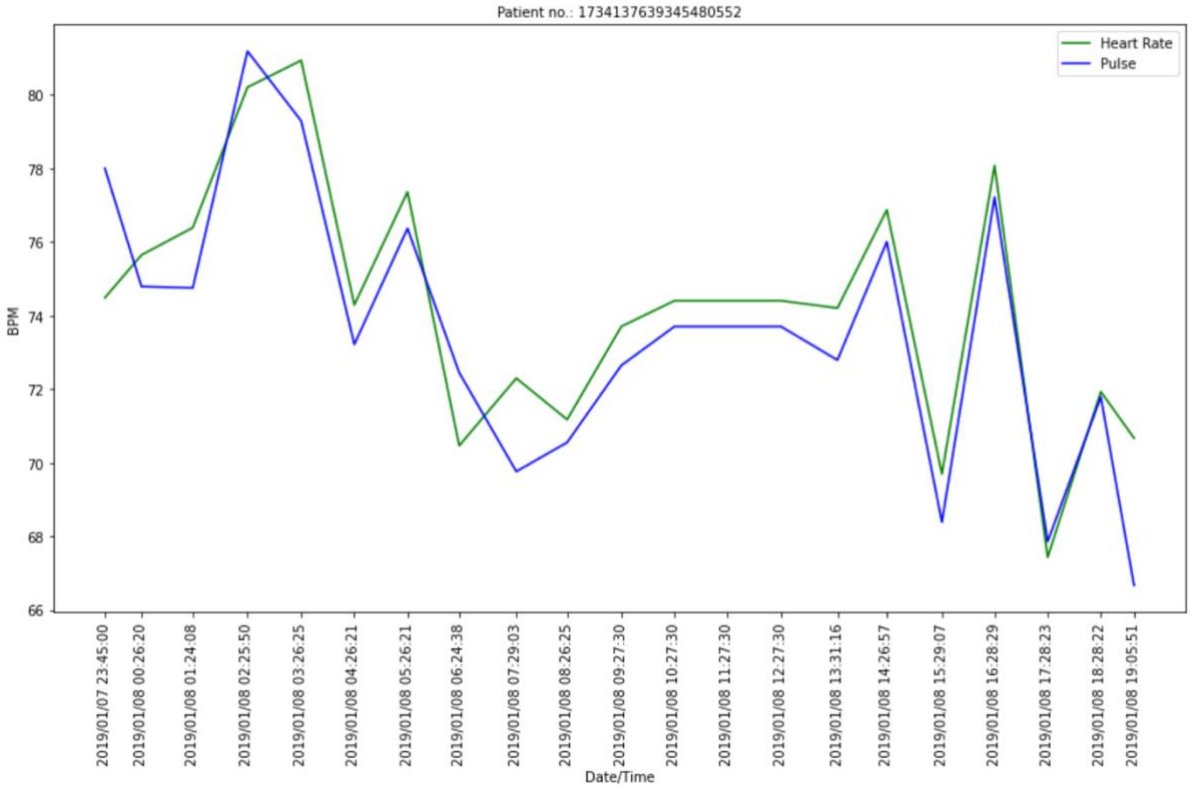

Figure 4.1 - Evolution of heart rate and pulse rate (BPM) over time

Figure 4.2 represents the output obtained from the system that provides us with information about Rita's variables (as mentioned, systolic and diastolic blood pressures values were NUUL).

```
The average heart rate is: 74.4
The average pulse is: 73.7
The average temperature is: 0.0
The average systolic pressure is: 0.0
The average diastolic pressure is: 0.0
```

Figure 4.2 - Average of different patient variables

Based on this calculation, the system will react whenever the value increases or decreases by 10%. The reaction will be to send an email to the nurses alerting them to the change in the patient's health status. If we have a change in Rita's heart rate, we will receive a message like the Figure 4.3.

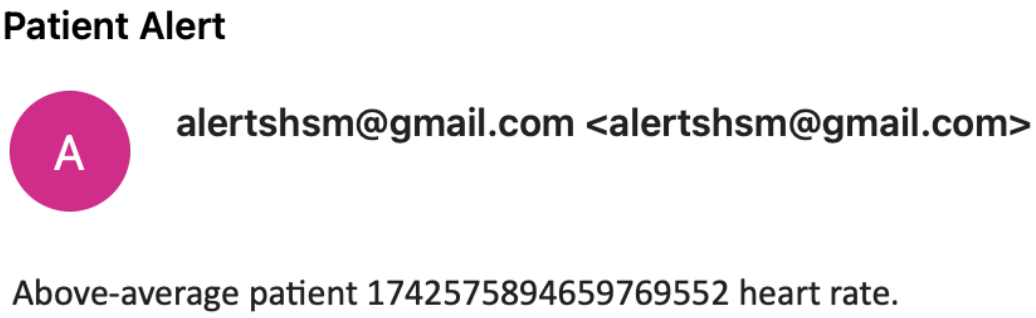

Figure 4.3 - Email alert about a change in the patient's heart rate

If it is the pulse, it will be as shown in Figure 4.4.

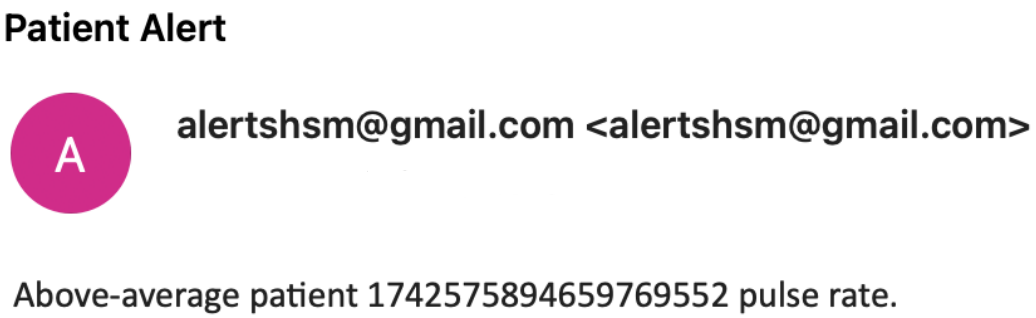

Figure 4.4 - Email alert about a change in the patient's pulse rate

4.3. System Evaluation

The whole project had the support of Prof. Dr. Luís Rosário, who pointed the direction that the system should take. In this sense, a presentation of the algorithm and the system was made, on 23rd of June of 2023, to health professionals in order to understand if the prototype met the needs that were mentioned in Chapter 1. Figure 4.1, shows the evaluation given by the experts. They were asked to answer the questions, indicating a number between 1 and 5, where 1 corresponds to I Do Not Agree and 5 corresponds to I Totally Agree.

Criteria	Question	Objective statement	#Eval 1	#Eval 2
Acceptance by Doctors	On a scale of 1-5, do you think it would have good acceptance among health professionals?	Possible acceptance among health care professionals	5	4
Ease of Use	On a scale of 1-5, how easy do you think it would be to use?	Simple and easy-to-read system	5	4
Clinical Impact	On a scale of 1-5, how would you rate the added value of the alerting system for clinical decision making?	Ability to decrease the response time in patients	4	5
Fit with organization	On a scale of 1-5, how do you think the alert system would fit into the current workflow?	A tool that would fit into the current workflow	4	5
Harnessing of recent technologies	On a scale of 1-5, do you think it would be a viable tool to help reduce hospital overload?	Feasibility to reduce hospital overload	5	4
Utility	On a scale of 1-5, how confident would you be in the alert system's ability to detect relevant changes in the patient's clinical status?	Confidence in the system as work tool	5	4

Table 4.1 - Results of the evaluation of the system by health professionals

Overall the evaluation is positive. According to the feedback that was transmitted during the meeting it was mentioned that initially the implementation of a new system could cause a certain entropy with the current workflow. However, over time it could be a promising tool.

5. Discussion and Conclusions

This thesis explored solutions to tackle our research question: "How can we use ML algorithms to accurately predict the risk of a patient exceeding the normal heart rate and develop an alert system, based on pseudo-anonymized clinical data from patients at Hospital Santa Maria, so that medical professionals can act quickly and efficiently?". To address this resarfch question, our proposed solutions are twofold:

- A machine learning model to accurately predict a patient's risk of exceeding the normal heart rate, trained, and tested on the HSM dataset, that was able to identify 1:15 hours in advance, the patients who were at a higher risk of exceeding the normal heart rate. The algorithm chosen was Logistics Regression (LR) showing a 91% recall.
- An alert system that works with our developed ML algorithm or with metaheuristics if no annotated data is available. The system has been shown to be able to provide immediate remote notification to healthcare professionals if a patient may exceed their heart rate 1:15 minutes in advance or in case of abnormalities according to specific metrics.

The potential benefits of this work are significant, as prevention and care for patients' heart health is crucial to reducing mortality and improving quality of life. This is because it can help detect problems early and allow timely medical interventions, avoiding unnecessary treatments and reducing costs to the healthcare system.

However, we recognize that the study has some limitations, such as the use of small a sample of the HSM dataset, with many null values in some variables, as well as and many text-based variables, not addressed in this study and currently being analyzed in other research activities being carried by our group.

In summary, we believe that this study can make a significant contribution to the field of medicine and technology by providing innovative solutions for prevention and care of patients' heart health.

Metadata

ADMISSIONAMBULATORYSTATUS

- The table consists of the status id, the admission id, and the ambulance status id.

ADMISSIONCHIEFCOMPLAINT

- This table has the admission id of the complaint, the id of the admission.

ADMISSIONDIAGNOSIS

- This table has the primary key, (dboid), the date of diagnosis, the severity of the patient as well as some additional information, some identifiers of the date of admission of the diagnosis in the record, comments on the diagnosis, the person responsible for entering the record, the priority of the record, whether the diagnosis is external or not.

ADMISSIONS

- This table contains the id of the admission, admid, when admitted to the hospital, preadmission, weight, height, age id of the patient, parents of the patient and when left the hospital.

ALLERGIES

- This table contains the id, description, and id of the allergy.

ALLERGYTYPES

- This table presents the id of the allergy type, description of it.

ANALYSES

- This table is composed of the primary key of the analyses, the test name, the index of the analysis, where it comes from, the foreign key that identifies the analysis and the id identifying the analysis type.

ATTENDINGTYPES

- This table has the id of who attended, the description of the same.

BLOODGROUPS

- This table has the id of the blood group, the description of it.

CATEGORIES

- This table is composed of the primary key of the table, the category description, the system code, the index category, and the key to the categorytype table.

CHIEFCOMPLAINTS

- This table is composed of the table id and the patient's chief complaint.

CNLCODES

- This table corresponds to codes from table RTDATA.

COUNTRIES

- This table consists of the country id, the description.

DEPARTMENTS

- It consists of the primary key , department description, the department code, the department type which is a foreign key to the departmentsypes table and the description of the department foreign key.

DIAGNOSES

- This table contains a diagnosis code and its description defined by the hospital.

DIAGNOSISTYPES

- This table represents the category of diagnosis defined by the hospital.

ENCNOTETYPENOTETYPEBLOCK

- This table corresponds to the annotation type.

ENCOUNTERNOTETYPE

- This table contains the id of the note type, the note id of the note on the patient.

ENCOUNTERS

- It gives us information about the event id identifier, the start and end, identifier for the picis date table, the event type.

ENVIRONMENTLOCATION

- It consists of a primary key, the start, the end, foreign key to the Environments table, foreign key to the Locations table, the status where -1- status unknown, 1- the admission has already been created but not yet started and 2 - the session has already started.

ENVIRONMENTMEDICALPROCEDURE

- **T**his table contains the id, foreign key of Environments, foreign key of Medical Procedure, performance, rank and staff, identification of the medical procedure, the event identifier that is associated and its description.

ENVIRONMENTS

- It has the primary key, the session start, end (start of the next session), picisdata foreign key, environment types, asatypes**.**

ENVIRONMENTTYPEGROUPS

- This table contains the primary key the group type name, the group type description, the process mode.

ENVIRONMENTTYPES

- This table contains the primary key the environment type, the name of the environment type, and the foreign key of Envrionment type grops.

ETHNICITIES

- It contains the primary key the patient's ethnicity and the patient's description.

EVENTDATA

- It is composed of the primary key, the date of the event that the record was added, remarks, update of the last record, Picis Data foreign key, event identifier, person responsible for the record, location of the environment reported from the event (-1), specification if it comes via another system, Sytem avent data foreign key.

EVENTS

- It is composed of a primary key, event description, use of event symbols (optional), whether the event appears in the SAM model and foreign key of Event type.

EVENTTYPES

- It is composed of a primary key, the key description, and points to the event type category.

FAMILIES

- It is composed of the primary key, the description of the species, the order they should be applied, foreign key of Categories, and foreign key of familybehavior.

FORMS

- It is composed of the primary key, the form, and the foreign key of formtypes.

LABRESULTS

- This table shows its primary key, the result value, lab number or text, if the lab result was edited (if so, it must be in the labresultauduted table), its matching values, labtests foreign key, partcomponent foreign key, labdatastatus foreign key, identification of the unit of measure of the result, the person responsible, remarks.

LABSOURCES

- Contains no information.

LABTESTS

- This table contains the respective primary key, the date the test was started, the date the test was completed, free text for comments, foreign key for the labsources table, foreign key for analyses, identifier of the person responsible for inserting the record and foreign key d picisdata.

LOCATIONS

- It is composed of a primary key, the bed identifier, the internal identifier, the name of the computer that is associated, a short version of the name, the location type (B - patient bed (from 'b'ed. Not connected to any computer); C - multi-bed workstation (from Computer. Not connected to any bed); D - bedside workstation (from 'd'ual.), hospital identifier, machine type.

MEDICALPROCEDURE

- This table is composed of the primary key, the description of the medical procedure, the procedure code, foreign key of the medicalproceduretypes table, the application version of the foreign key and indication of the medical procedure.

MEDICALPROCEDURETYPES

- It is composed of the primary key, the type of medical procedure.

MEDICATIONS

- It is composed of the primary key, the description of the medication, comments about the medication and the id of the generic.

NOTETYPEBLOCKS

- This table is composed of a primary key, the name, the short name of the note type and the foreign key for the note types of table.

NOTETYPES

- It is composed of the id of the type of note, its description, its index.

ORDERS

- It contains the primary key, name of the order, when it was created, free text to explain the conditions and the order of administration, what time it started, what time it finished, the indicator and the maximum dose.

ORDERTASKSTATUS

- It contains the id, the description of it.

PARTCOMPONENT

- This table is composed of the primary key, the Parts foreign key and the Components foreign key.

PARTS

- It is composed of the table id, its description, and the analysis id.

PATIENTALLERGY

- This table is composed of table id, when the patient entered, and left the hospital allergy index, comments, patient id, allergy id and note associated with the patient.

PATIENTPRECAUTION

- This table is composed of table id, when the patient entered and left the hospital, patient id, patient precautions id, and comments.

PATIENTS

- The patient table has the corresponding id for each patient, birthday date, blood group sex and ethnicity.

PSICADATA

- Contains the patient id, admission id.

PRECAUTIONS

- It has the precaution id and its description, the precaution type id.

PRECAUTIONTYPES

- It consists of the precaution type, the description.

ROUTES

- This table is made up of the table id, its description, and the id of the type of route the patient received treatment.

RTDATA

- Table composed by id of RTDATA, when the patient was admitted to the hospital, the last update, and other variables that are codes.

SCHEDULES

- It corresponds to the id of the table its description the frequency and duration of order schedules that can be selected when creating standard orders or when prescribing custom orders.

SCOREGROUPS

- It consists of id, description, index (ranging from 0 to 22) and the treatment id.

SCOREITEMS

- This table consists of the table id, its description, index and the patient's respective pain scale.

SELEC

- It shows the status id, the description.

STAFF

- This table contains the table id, the first and last name, the password of the same.

STAFFATTENDINGTYPE

- This table is composed of the table id, the foreign key of the staff table and the foreign key of the ATTTYP table.

STAFFTYPES

- This table is composed of the primary key of the table, its description, and the type of staff.

TABLES

- Unformed table, impossible to analyse.

TASK

- Information from the beginning of the task creation.

TREATMENTS

- It consists of the treatment id, the generic name, the business name, the family id (foreign key of the family table), the group id and group validation id.

UNITS

- This table consists of its id, symbol description, conversion factor, unit type, numerator id, and weight id.

SYSDIAGRAMS

- Table cannot be parsed.

References

[1] "Cardiovascular diseases (CVDs)." https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Sep. 26, 2022).

[2] "Cardiovascular diseases - PAHO/WHO | Pan American Health Organization." https://www.paho.org/en/topics/cardiovascular-diseases (accessed Sep. 26, 2022).

[3] "Cardiovascular diseases - PAHO/WHO | Pan American Health Organization." https://www.paho.org/en/topics/cardiovascular-diseases (accessed Sep. 26, 2022).

[4] "Cardiovascular diseases - PAHO/WHO | Pan American Health Organization." https://www.paho.org/en/topics/cardiovascular-diseases (accessed Jan. 23, 2023).

[5] "Implementation roadmap 2023-2030 for the Global action plan." https://www.who.int/teams/noncommunicable-diseases/governance/roadmap (accessed Mar. 17, 2023).

[6] "World Health Assembly approves a global implementation roadmap to accelerate action on noncommunicable diseases (NCDs)." https://www.who.int/news-room/feature-stories/detail/world-health-assembly-approves-a-global-implementation-roadmap-to-accelerate-action-on-noncommunicable-diseases-(ncds) (accessed Feb. 24, 2023).

[7] F. Custodis, J.-C. Reil, U. Laufs, and M. Böhm, "Heart rate: A global target for cardiovascular disease and therapy along the cardiovascular disease continuum," J. Cardiol., vol. 62, no. 3, pp. 183–187, Sep. 2013, doi: 10.1016/j.jjcc.2013.02.018.

[8] Å. Hjalmarson, "Heart rate: an independent risk factor in cardiovascular disease," Eur. Heart J. Suppl., vol. 9, no. suppl_F, pp. F3–F7, Sep. 2007, doi: 10.1093/eurheartj/sum030.

[9] "What is Big Data Analytics and Why is it Important?," Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics (accessed Jan. 23, 2023).

[10] "Typical Application of Big Data Technology in Different Fields," J. Res. Sci. Eng., vol. 4, no. 9, Sep. 2022, doi: 10.53469/jrse.2022.04(09).01.

[11] "How to manage complexity and realize the value of big data," Smarter Business Review, May 28, 2020. https://www.ibm.com/blogs/services/2020/05/28/how-to-manage-complexity-and-realize-the-value-of-big-data/ (accessed Jan. 25, 2023).

[12] I. K. Nti, J. A. Quarcoo, J. Aning, and G. K. Fosu, "A mini-review of machine learning in big data analytics: Applications, challenges, and prospects," Big Data Min. Anal., vol. 5, no. 2, pp. 81–97, Jun. 2022, doi: 10.26599/BDMA.2021.9020028.

[13] D. T. Huff, A. J. Weisman, and R. Jeraj, "Interpretation and visualization techniques for deep learning models in medical imaging," Phys. Med. Biol., vol. 66, no. 4, p. 04TR01, Feb. 2021, doi: 10.1088/1361-6560/abcd17.

[14] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," in 2012 Japan-Egypt Conference on Electronics, Communications and Computers, Mar. 2012, pp. 173–177. doi: 10.1109/JEC-ECC.2012.6186978.

[15] R. J. Oskouei, N. M. Kor, and S. A. Maleki, "Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges," Am. J. Cancer Res., vol. 7, no. 3, pp. 610–627, Mar. 2017.

[16] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," Int. J. Comput. Appl., vol. 17, no. 8, pp. 43–48, Mar. 2011, doi: 10.5120/2237-2860.

[17] "Kidney Disease as a Risk Factor for Development of Cardiovascular Disease." https://www.ahajournals.org/doi/epub/10.1161/01.CIR.0000095676.90936.80 (accessed Jan. 23, 2023).

[18] R. Pastorino, C. De Vito, G. Migliara, K. Glocker, I. Binenbaum, W. Ricciardi, and S. Boccia, "Benefits and challenges of Big Data in healthcare: an overview of the European initiatives," Eur. J. Public Health, vol. 29, no. Supplement_3, pp. 23–27, Oct. 2019, doi: 10.1093/eurpub/ckz168.
[19] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.
[20] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," Mach. Learn. Knowl. Extr., vol. 3, no. 2, Art. no. 2, Jun. 2021, doi: 10.3390/make3020020.
[21] W. Zeng, Z. Lin, C. Yuan, Q. Wang, F. Liu, and Y. Wang, "Detection of heart valve disorders from PCG signals using TQWT, FA-MVEMD, Shannon energy envelope and deterministic learning," Artif. Intell. Rev., vol. 54, no. 8, pp. 6063–6100, Dec. 2021, doi: 10.1007/s10462-021-09969-z.
[22] W. Wei, G. Zhan, X. Wang, P. Zhang, and Y. Yan, "A novel method for automatic heart murmur diagnosis using phonocardiogram," presented at the ACM International Conference Proceeding Series, 2019. doi: 10.1145/3358331.3358368.
[23] K. Suhas, R. H. Kumar, S. H. Nayak, and B. N. Krupa, "A hybrid model for recognizing cardiac murmurs from phonocardiogram signal," presented at the 2016 IEEE Annual India Conference, INDICON 2016, 2017. doi: 10.1109/INDICON.2016.7839002.
[24] G. Strange, D. S. Celermajer, T. Marwick, D. Prior, M. Ilton, J. Codde, G. M. Scalia, S. Stewart, M. Bulsara, E. Gabbay, and D. Playford, "The National Echocardiography Database Australia (NEDA): Rationale and methodology," Am. Heart J., vol. 204, pp. 186–189, 2018, doi: 10.1016/j.ahj.2018.07.001.
[25] W. R. Thompson, A. J. Reinisch, M. J. Unterberger, and A. J. Schriefl, "Artificial Intelligence-Assisted Auscultation of Heart Murmurs: Validation by Virtual Clinical Trial," Pediatr. Cardiol., vol. 40, no. 3, pp. 623–629, 2019, doi: 10.1007/s00246-018-2036-z.
[26] J. D. Thomas, O. M. Petrescu, S. K. Moualla, M. Dobbles, J. C. Hays, E. Rodriguez, and G. R. Barnhart, "Artificial intelligence to assist physicians in identifying patients with severe aortic stenosis," Intell.-Based Med., vol. 6, 2022, doi: 10.1016/j.ibmed.2022.100059.
[27] S. R. Thiyagaraja, R. Dantu, P. L. Shrestha, A. Chitnis, M. A. Thompson, P. T. Anumandla, T. Sarma, and S. Dantu, "A novel heart-mobile interface for detection and classification of heart sounds," Biomed. Signal Process. Control, vol. 45, pp. 313–324, 2018, doi: 10.1016/j.bspc.2018.05.008.
[28] A. Shokouhmand, N. Aranoff, E. Driggin, P. Green, and N. Tavassolian, "Efficient detection of aortic stenosis using morphological characteristics of cardiomechanical signals and heart rate variability parameters," Sci. Rep., vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-03441-2.
[29] Z. Shi, N. Dai, R. Liu, J. Wang, S. Cai, and N. Hu, "Machine-Learning-based Aortic Stenosis Detection for Electronic Stethoscope," presented at the 13th International Conference on Wireless Communications and Signal Processing, WCSP 2021, 2021. doi: 10.1109/WCSP52459.2021.9613207.
[30] N. G. Kang, Y. J. Suh, K. Han, Y. J. Kim, and B. W. Choi, "Performance of prediction models for diagnosing severe aortic stenosis based on aortic valve calcium on cardiac

computed tomography: Incorporation of radiomics and machine learning," Korean J. Radiol., vol. 22, no. 3, pp. 334–343, 2021, doi: 10.3348/kjr.2020.0099.
[31] S. Goto, K. Mahara, L. Beussink-Nelson, H. Ikura, Y. Katsumata, J. Endo, H. Gaggin, S. Shah, Y. Itabashi, C. MacRae, and R. Deo, "Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms," Nat. Commun., vol. 12, no. 1, May 2021, doi: 10.1038/s41467-021-22877-8.
[32] S. K. Ghosh, R. N. Ponnalagu, R. K. Tripathy, and U. R. Acharya, "Deep Layer Kernel Sparse Representation Network for the Detection of Heart Valve Ailments from the Time-Frequency Representation of PCG Recordings," BioMed Res. Int., vol. 2020, 2020, doi: 10.1155/2020/8843963.
[33] S. Calcagno, G. Biondizoccai, T. Stankovic, E. Szabo, A. B. Szabo, and I. Kecskes, "Novel tech throws knock-out punch to ECG improving GP referral decisions to cardiology," Open Heart, vol. 9, no. 1, 2022, doi: 10.1136/openhrt-2021-001852.
[34] P. Garcia-Pavia, C. Rapezzi, Y. Adler, M. Arad, C. Basso, A. Brucato, I. Burazor, A. L. P. Caforio, T. Damy, U. Eriksson, M. Fontana, J. D. Gillmore, E. Gonzalez-Lopez, M. Grogan, S. Heymans, M. Imazio, I. Kindermann, A. V. Kristen, M. S. Maurer, G. Merlini, A. Pantazis, S. Pankuweit, A. G. Rigopoulos, and A. Linhart, "Diagnosis and treatment of cardiac amyloidosis. A position statement of the European Society of Cardiology Working Group on Myocardial and Pericardial Diseases," Eur. J. Heart Fail., vol. 23, no. 4, pp. 512–526, 2021, doi: 10.1002/ejhf.2140.
[35] H.-G. Jeong, B. J. Kim, T. Kim, J. Kang, J. Y. Kim, J. Kim, J.-T. Kim, J.-M. Park, J. G. Kim, J.-H. Hong, K. B. Lee, T. H. Park, D.-H. Kim, C. W. Oh, M.-K. Han, and H.-J. Bae, "Classification of cardioembolic stroke based on a deep neural network using chest radiographs," EBioMedicine, vol. 69, 2021, doi: 10.1016/j.ebiom.2021.103466.
[36] A. M. Alqudah, H. Alquran, and I. A. Qasmieh, "Classification of heart sound short records using bispectrum analysis approach images and deep learning," Netw. Model. Anal. Health Inform. Bioinforma., vol. 9, no. 1, 2020, doi: 10.1007/s13721-020-00272-5.
[37] M. Cohen-Shelly, Z. I. Attia, P. A. Friedman, S. Ito, B. A. Essayagh, W.-Y. Ko, D. H. Murphree, H. I. Michelena, M. Enriquez-Sarano, R. E. Carter, P. W. Johnson, P. A. Noseworthy, F. Lopez-Jimenez, and J. K. Oh, "Electrocardiogram screening for aortic valve stenosis using artificial intelligence," Eur. Heart J., vol. 42, no. 30, pp. 2885–2896, 2021, doi: 10.1093/eurheartj/ehab153.
[38] M. Wang, B. Guo, Y. Hu, Z. Zhao, C. Liu, and H. Tang, "Transfer Learning Models for Detecting Six Categories of Phonocardiogram Recordings," J. Cardiovasc. Dev. Dis., vol. 9, no. 3, Mar. 2022, doi: 10.3390/jcdd9030086.
[39] W. Zeng, B. Su, C. Yuan, and Y. Chen, "Automatic detection of heart valve disorders using Teager-Kaiser energy operator, rational-dilation wavelet transform and convolutional neural networks with PCG signals," Artif. Intell. Rev., doi: 10.1007/s10462-022-10184-7.
[40] J. S. Chorba, A. M. Shapiro, L. Le, J. Maidens, J. Prince, S. Pham, M. M. Kanzawa, D. N. Barbosa, C. Currie, C. Brooks, B. E. White, A. Huskin, J. Paek, J. Geocaris, D. Elnathan, R. Ronquillo, R. Kim, Z. H. Alam, V. S. Mahadevan, S. G. Fuller, G. W. Stalker, S. A. Bravo, D. Jean, J. J. Lee, M. Gjergjindreaj, C. G. Mihos, S. T. Forman, S. Venkatraman, P. M. McCarthy, and J. D. Thomas, "Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform," J. Am. Heart Assoc., vol. 10, no. 9, 2021, doi: 10.1161/JAHA.120.019905.
[41] A. Budai, F. Suhai, K. Csorba, Z. Dohy, L. Szabo, B. Merkely, and H. Vago, "Automated Classification of Left Ventricular Hypertrophy on Cardiac MRI," Appl. Sci.-BASEL, vol. 12, no. 9, May 2022, doi: 10.3390/app12094151.

[42] S. Feng, X. Wu, A. Bao, G. Lin, P. Sun, H. Cen, S. Chen, Y. Liu, W. He, Z. Pang, and H. Zhang, "Machine learning-aided detection of heart failure (LVEF ≤ 49%) by using ballistocardiography and respiratory effort signals," Front. Physiol., vol. 13, 2023, doi: 10.3389/fphys.2022.1068824.
[43] Q. Su, Q. Liu, R. I. Lau, J. Zhang, Z. Xu, Y. K. Yeoh, T. W. H. Leung, W. Tang, L. Zhang, J. Q. Y. Liang, Y. K. Yau, J. Zheng, C. Liu, M. Zhang, C. P. Cheung, J. Y. L. Ching, H. M. Tun, J. Yu, F. K. L. Chan, and S. C. Ng, "Faecal microbiome-based machine learning for multi-class disease diagnosis," Nat. Commun., vol. 13, no. 1, 2022, doi: 10.1038/s41467-022-34405-3.
[44] M. Bucholc, D. Bradley, D. Bennett, L. Patterson, R. Spiers, D. Gibson, H. Van Woerden, and A. J. Bjourson, "Identifying pre-existing conditions and multimorbidity patterns associated with in-hospital mortality in patients with COVID-19," Sci. Rep., vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-20176-w.
[45] A. Jahangirimehr, E. Abdolahi Shahvali, S. M. Rezaeijo, A. Khalighi, A. Honarmandpour, F. Honarmandpour, M. Labibzadeh, N. Bahmanyari, and S. Heydarheydari, "Machine learning approach for automated predicting of COVID-19 severity based on clinical and paraclinical characteristics: Serum levels of zinc, calcium, and vitamin D," Clin. Nutr. ESPEN, vol. 51, pp. 404–411, 2022, doi: 10.1016/j.clnesp.2022.07.011.
[46] M. P. McRae, K. S. Rajsri, T. M. Alcorn, and J. T. McDevitt, "Smart Diagnostics: Combining Artificial Intelligence and In Vitro Diagnostics," Sensors, vol. 22, no. 17, 2022, doi: 10.3390/s22176355.
[47] O. Attallah, "An Intelligent ECG-Based Tool for Diagnosing COVID-19 via Ensemble Deep Learning Techniques," Biosensors, vol. 12, no. 5, 2022, doi: 10.3390/bios12050299.
[48] A. Romaszko-Wojtowicz, S. Maksymowicz, A. Jarynowski, Ł. Jaśkiewicz, Ł. Czekaj, and A. Doboszyńska, "Telemonitoring in Long-COVID Patients—Preliminary Findings," Int. J. Environ. Res. Public. Health, vol. 19, no. 9, 2022, doi: 10.3390/ijerph19095268.
[49] J. Ahamed, R. N. Mir, and M. A. Chishti, "Industry 4.0 oriented predictive analytics of cardiovascular diseases using machine learning, hyperparameter tuning and ensemble techniques," Ind. Robot, vol. 49, no. 3, pp. 544–554, 2022, doi: 10.1108/IR-10-2021-0240.
[50] S. Kulbayeva, K. Tazhibayeva, L. Seiduanova, I. Smagulova, A. Mussina, S. Tanabayeva, I. Fakhradiyev, and T. Saliev, "The Recent Advances of Mobile Healthcare in Cardiology Practice," Acta Inform. Medica, vol. 30, no. 3, pp. 236–250, 2022, doi: 10.5455/aim.2022.30.236-250.
[51] T. Sadad, S. A. C. Bukhari, A. Munir, A. Ghani, A. M. El-Sherbeeny, and H. T. Rauf, "Detection of Cardiovascular Disease Based on PPG Signals Using Machine Learning with Cloud Computing," Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/1672677.
[52] S. Diwan, G. S. Thakur, S. K. Sahu, M. Sahu, and N. K. Swamy, "Predicting Heart Diseases through Feature Selection and Ensemble Classifiers," presented at the Journal of Physics: Conference Series, 2022. doi: 10.1088/1742-6596/2273/1/012027.
[53] Y. L. Kamala, K. V. S. N. R. Rao, and B. M. Josephine, "Comparison and Evaluation of Studies on Precision Medicine using AI," presented at the International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2022 - Proceedings, 2022, pp. 330–335. doi: 10.1109/ICSCDS53736.2022.9760969.
[54] P. Pournazari, A. L. Spangler, F. Ameer, K. K. Hagan, M. E. Tano, M. Chamsi-Pasha, L. H. Chebrolu, W. A. Zoghbi, K. Nasir, and S. F. Nagueh, "Cardiac involvement in hospitalized patients with COVID-19 and its incremental value in outcomes prediction," Sci. Rep., vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-98773-4.
[55] A. Tariq, L. A. Celi, J. M. Newsome, S. Purkayastha, N. K. Bhatia, H. Trivedi, J. W. Gichoya, and I. Banerjee, "Patient-specific COVID-19 resource utilization prediction

using fusion AI model," Npj Digit. Med., vol. 4, no. 1, 2021, doi: 10.1038/s41746-021-00461-0.
[56] S. Munjral, P. Ahluwalia, A. D. Jamthikar, A. Puvvula, L. Saba, G. Faa, I. M. Singh, P. S. Chadha, M. Turk, A. M. Johri, N. N. Khanna, K. Viskovic, S. Mavrogeni, J. R. Laird, G. Pareek, M. Miner, D. W. Sobel, A. Balestrieri, P. P. Sfikakis, G. Tsoulfas, A. Protogerou, P. Misra, V. Agarwal, G. D. Kitas, R. Kolluri, J. Teji, M. Al-Maini, S. K. Dhanjil, M. Sockalingam, A. Saxena, A. Sharma, V. Rathore, M. Fatemi, A. Alizad, V. Viswanathan, P. K. Krishnan, T. Omerzu, S. Naidu, A. Nicolaides, and J. S. Suri, "Nutrition, atherosclerosis, arterial imaging, cardiovascular risk stratification, and manifestations in COVID-19 framework: a narrative review," Front. Biosci. - Landmark, vol. 26, no. 11, pp. 1312–1339, 2021, doi: 10.52586/5026.
[57] O. Krysko, E. Kondakova, O. Vershinina, E. Galova, A. Blagonravova, E. Gorshkova, C. Bachert, M. Ivanchenko, D. V. Krysko, and M. Vedunova, "Artificial Intelligence Predicts Severity of COVID-19 Based on Correlation of Exaggerated Monocyte Activation, Excessive Organ Damage and Hyperinflammatory Syndrome: A Prospective Clinical Study," Front. Immunol., vol. 12, 2021, doi: 10.3389/fimmu.2021.715072.
[58] P. Mohandas, P. R. Aswin, A. John, M. Madhu, G. Thomas, and V. Kurupath, "Automated cardiac condition diagnosis using AI based ECG analysis system for school children," presented at the 2021 8th International Conference on Smart Computing and Communications: Artificial Intelligence, AI Driven Applications for a Smart World, ICSCC 2021, 2021, pp. 362–366. doi: 10.1109/ICSCC51209.2021.9528098.
[59] Y. Chen, L. Ouyang, F. S. Bao, Q. Li, L. Han, H. Zhang, B. Zhu, Y. Ge, P. Robinson, M. Xu, J. Liu, and S. Chen, "A multimodality machine learning approach to differentiate severe and nonsevere COVID-19: Model development and validation," J. Med. Internet Res., vol. 23, no. 4, 2021, doi: 10.2196/23948.
[60] L. Abuabara, M. G. Valeriano, C. R. Veiga Kiffer, H. H. Yanasse, and A. C. Lorena, "Using Machine Learning to support health system planning during the COVID-19 pandemic: a case study using data from São José dos Campos (Brazil)," CLEI Eletronic J. CLEIej, vol. 24, no. 3, pp. 1–21, 2021, doi: 10.19153/CLEIEJ.24.3.5.
[61] A. Haleem, M. Javaid, R. P. Singh, and R. Suman, "Applications of Artificial Intelligence (AI) for cardiology during COVID-19 pandemic," Sustain. Oper. Comput., vol. 2, pp. 71–78, 2021, doi: 10.1016/j.susoc.2021.04.003.
[62] P. S. Lee, S. Koo, and S. Panter, "The value of physical examination in the era of telemedicine," J. R. Coll. Physicians Edinb., vol. 51, no. 1, pp. 85–90, 2021, doi: 10.4997/JRCPE.2021.122.
[63] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice," Digit. Biomark., vol. 5, no. 1, pp. 78–88, 2021, doi: 10.1159/000515346.
[64] Y. Fu, Y. Zhang, and B. L. Khoo, "Liquid biopsy technologies for hematological diseases," Med. Res. Rev., vol. 41, no. 1, pp. 246–274, 2021, doi: 10.1002/med.21731.
[65] C. D'Ambrosia, H. Christensen, and E. Aronoff-Spencer, "Computing SARS-CoV-2 infection risk from symptoms, imaging, and test data: Diagnostic model development," J. Med. Internet Res., vol. 22, no. 12, 2020, doi: 10.2196/24478.
[66] A. Zimmerman and D. Kalra, "Usefulness of machine learning in COVID-19 for the detection and prognosis of cardiovascular complications," Rev. Cardiovasc. Med., vol. 21, no. 3, pp. 345–352, 2020, doi: 10.31083/j.rcm.2020.03.120.
[67] M. K. Chaitanya, L. D. Sharma, J. Rahul, D. Sharma, and A. Roy, "Artificial intelligence based approach for categorization of COVID-19 ECG images in presence of other cardiovascular disorders," Biomed. Phys. Eng. Express, vol. 9, no. 3, 2023, doi: 10.1088/2057-1976/acbd53.

[68] T. Rahman, A. Akinbi, M. E. H. Chowdhury, T. A. Rashid, A. Şengür, A. Khandakar, K. R. Islam, and A. M. Ismael, "COV-ECGNET: COVID-19 detection using ECG trace images with deep convolutional neural network," Health Inf. Sci. Syst., vol. 10, no. 1, 2022, doi: 10.1007/s13755-021-00169-1.
[69] A. Agrawal, A. Chauhan, M. K. Shetty, G. M. P, M. D. Gupta, and A. Gupta, "ECG-iCOVIDNet: Interpretable AI model to identify changes in the ECG signals of post-COVID subjects," Comput. Biol. Med., vol. 146, 2022, doi: 10.1016/j.compbiomed.2022.105540.
[70] E. Irmak, "COVID-19 disease diagnosis from paper-based ECG trace image data using a novel convolutional neural network model," Phys. Eng. Sci. Med., vol. 45, no. 1, pp. 167–179, 2022, doi: 10.1007/s13246-022-01102-w.
[71] N. Ji, T. Xiang, P. Bonato, N. H. Lovell, S.-Y. Ooi, D. A. Clifton, M. Akay, X.-R. Ding, B. P. Yan, V. Mok, D. I. Fotiadis, and Y.-T. Zhang, "Recommendation to Use Wearable-Based mHealth in Closed-Loop Management of Acute Cardiovascular Disease Patients during the COVID-19 Pandemic," IEEE J. Biomed. Health Inform., vol. 25, no. 4, pp. 903–908, 2021, doi: 10.1109/JBHI.2021.3059883.
[72] B. T. Thomas, N. J. Binu, A. Varghese, A. A. Peter, G. Thomas, and V. Kurupath, "Artificial Intelligence Based Phonocardiogram Proposed for Cardiac Screening of School Children," presented at the International Conference on Futuristic Technologies in Control Systems and Renewable Energy, ICFCR 2020, 2020. doi: 10.1109/ICFCR50903.2020.9249983.
[73] E. A. Takahashi, L. H. Schwamm, O. M. Adeoye, O. Alabi, E. Jahangir, S. Misra, and C. H. Still, "An Overview of Telehealth in the Management of Cardiovascular Disease: A Scientific Statement from the American Heart Association," Circulation, vol. 146, no. 25, pp. E558–E568, 2022, doi: 10.1161/CIR.0000000000001107.
[74] O. Irtyuga, G. Kopanitsa, A. Kostareva, O. Metsker, V. Uspensky, G. Mikhail, G. Faggian, G. Sefieva, I. Derevitskii, A. Malashicheva, and E. Shlyakhto, "Application of Machine Learning Methods to Analyze Occurrence and Clinical Features of Ascending Aortic Dilatation in Patients with and without Bicuspid Aortic Valve," J. Pers. Med., vol. 12, no. 5, May 2022, doi: 10.3390/jpm12050794.
[75] X. Diao, Y. Huo, Z. Yan, H. Wang, J. Yuan, Y. Wang, J. Cai, and W. Zhao, "An Application of Machine Learning to Etiological Diagnosis of Secondary Hypertension: Retrospective Study Using Electronic Medical Records," JMIR Med. Inform., vol. 9, no. 1, Jan. 2021, doi: 10.2196/19739.
[76] S. Críales-Vera, H. Saucedo-Orozco, P. Iturralde-Torres, G. Martínez-Mota, E. Dávila-Medina, V. Guarner-Lans, L. Manzano-Pech, I. Pérez-Torres, and M. E. Soto, "Tomography and Prognostic Indices in the State of the Art of Evaluation in Hospitalized Patients with COVID-19 Pneumonia," Pathogens, vol. 11, no. 11, 2022, doi: 10.3390/pathogens11111281.
[77] M. Cortes, "Breve olhar sobre o estado da saúde em Portugal," Sociol. Probl. E Práticas, no. 80, Art. no. 80, Jan. 2016.
[78] J. M. Meza, M. Slieker, E. H. Blackstone, L. Mertens, W. M. DeCampli, J. K. Kirklin, M. Karimi, P. Eghtesady, K. Pourmoghadam, R. W. Kim, P. T. Burch, M. L. Jacobs, T. Karamlou, B. W. McCrindle, and Congenital Heart Surgeons' Society, "A novel, data-driven conceptualization for critical left heart obstruction," Comput. Methods Programs Biomed., vol. 165, pp. 107–116, 2018, doi: 10.1016/j.cmpb.2018.08.014.
[79] "Monitorização Mensal Hospitais - Capacidade Utilizada." https://benchmarking-acss.min-saude.pt/MH_CapacidadeUtilizadaDashboard (accessed Feb. 21, 2023).
[80] J. Perktold, Skipper Seabold, K. Sheppard, ChadFulton, Kerby Shedden, Jbrockmendel, J-Grana6, P. Quackenbush, V. Arel-Bundock, W. McKinney, I. Langmore, B. Baker, R.

Gommers, Yogabonito, S-Scherrer, E. Zhurko, M. Brett, E. Giampieri, Yichuan Liu, J. Millman, P. Hobson, Vincent, P. Roy, T. Augspurger, Tvanzyl, Alexbrc, T. Hartley, F. Perez, Y. Tamiya, and Y. Halchenko, "statsmodels/statsmodels: Release 0.14.0." Zenodo, May 05, 2023. doi: 10.5281/ZENODO.593847.
[81] "Getting Started," scikit-learn. https://scikit-learn/stable/getting_started.html (accessed May 13, 2023).
[82] "1.9. Naive Bayes," scikit-learn. https://scikit-learn/stable/modules/naive_bayes.html (accessed May 13, 2023).
[83] "sklearn.linear_model.LogisticRegression," scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed May 13, 2023).
[84] "1.6. Nearest Neighbors," scikit-learn. https://scikit-learn/stable/modules/neighbors.html (accessed May 13, 2023).
[85] "sklearn.ensemble.RandomForestClassifier," scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed May 13, 2023).
[86] "pandas.DataFrame — pandas 2.0.1 documentation." https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html (accessed May 13, 2023).
[87] "NumPy documentation — NumPy v1.24 Manual." https://numpy.org/doc/stable/ (accessed May 13, 2023).
[88] "smtplib — SMTP protocol client," Python documentation. https://docs.python.org/3/library/smtplib.html (accessed May 13, 2023).
[89] "threading — Thread-based parallelism," Python documentation. https://docs.python.org/3/library/threading.html (accessed May 13, 2023).
[90] Andrei-Popov, "Transport Layer Security protocol," Feb. 14, 2023. https://learn.microsoft.com/en-us/windows-server/security/tls/transport-layer-security-protocol (accessed May 13, 2023).
[91] "Secure Socket Layer (SSL)," GeeksforGeeks, Jun. 10, 2019. https://www.geeksforgeeks.org/secure-socket-layer-ssl/ (accessed May 13, 2023).