UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA BIOMÉDIA

TRABAJO FIN DE GRADO

DESIGN AND DEVELOPMENT OF A MACHINE LEARNING SYSTEM FOR PREDICTING PATIENT ILLNESSES BASED ON DEMOGRAPHIC AND LABORATORY DATA

MATTEO TARONNA NESPECA
JUNIO 2023

TRABAJO DE FIN DE GRADO

Título:	Diseño y Desarrollo de un Sistema de Aprendizaje Automático para la Predicción de Enfermedades de Pacientes basado en Datos Demográficos y de Laboratorio
Título (inglés):	Design and Development of a Machine Learning System for Predicting Patient Illnesses based on Demographic and Laboratory Data
Autor:	MATTEO TARONNA NESPECA
Tutor:	OSCAR ARAQUE
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

DESIGN AND DEVELOPMENT OF A MACHINE LEARNING SYSTEM FOR PREDICTING PATIENT ILLNESSES BASED ON DEMOGRAPHIC AND LABORATORY DATA

MATTEO TARONNA NESPECA

Junio 2023

Resumen

En los últimos años, los avances en inteligencia artificial y big data han abierto nuevas posibilidades en el campo de la salud y la medicina. Sin embargo, es necesario realizar más investigaciones para garantizar la precisión y confiabilidad de los modelos predictivos en este ámbito. Este estudio tiene como objetivo cerrar esta brecha explorando la aplicación de algoritmos de aprendizaje automático para predecir enfermedades utilizando información demográfica y resultados de laboratorio. El objetivo principal es desarrollar un modelo altamente preciso que pueda diagnosticar enfermedades y proporcionar enfoques de atención médica personalizados.

Para lograr este objetivo, se utilizaron diversas fuentes de datos, incluido el conjunto de datos EMRBots, que sirve como base para el entrenamiento y la evaluación. Además de EMRBots, se incorporaron dos repositorios especializados centrados en la detección de enfermedades cardíacas en la investigación. Estos conjuntos de datos proporcionaron conocimientos valiosos sobre la salud cardiovascular y mejoraron aún más la precisión del modelo predictivo.

A lo largo del estudio, se implementó una metodología integral. Se emplearon técnicas de preprocesamiento de datos para limpiar y formatear los conjuntos de datos, asegurando una calidad y consistencia óptimas. Se utilizaron varios algoritmos de aprendizaje automático, respaldados por la potente biblioteca de análisis de datos Pandas, la reconocida biblioteca de aprendizaje automático scikit-learn y la versátil biblioteca de cómputo numérico NumPy, para entrenar y ajustar el modelo predictivo.

El rendimiento del sistema desarrollado se evaluó minuciosamente utilizando métricas adecuadas, como precisión, recall y puntuación F1. Al integrar diversos conjuntos de datos, emplear tecnologías de vanguardia y centrarse en la detección de enfermedades cardíacas, este estudio ofrece conocimientos valiosos sobre el desarrollo de un sistema de aprendizaje automático sólido y preciso para la predicción de enfermedades.

Palabras clave: Inteligencia artificial, Big Data, Enfermedades cardíacas, EMRBots, Preprocesamiento de datos, Algoritmos de aprendizaje automático, Pandas, Scikit-Learn, NumPy, Exactitud, Precisión, Recall, Puntuación F1

Abstract

In recent years, advancements in artificial intelligence and big data have opened up new possibilities for healthcare and medicine. However, there is a need for further research to ensure the accuracy and reliability of predictive models in this domain. This study aims to bridge this gap by exploring the application of machine learning algorithms to predict diseases using demographic information and laboratory results. The primary objective is to develop a highly accurate model that can diagnose illnesses and provide personalized healthcare approaches.

To achieve this objective, an extensive range of data sources was utilized, including the EMRBots dataset, which serves as the foundation for training and evaluation. In addition to EMRBots, two specialized repositories focusing on cardiac disease detection were incorporated into the research. These datasets provided valuable insights into cardiovascular health and further enhanced the accuracy of the predictive model.

Throughout the study, a comprehensive methodology was implemented. Data preprocessing techniques were employed to clean and format the datasets, ensuring optimal quality and consistency. Various machine learning algorithms, supported by the powerful data analysis library Pandas, the renowned machine learning library scikit-learn, and the versatile numerical computing library NumPy, were utilized to train and fine-tune the predictive model.

The performance of the developed system was thoroughly evaluated using appropriate metrics, assessing its accuracy, precision, recall, and F1 score. By integrating various datasets, employing cutting-edge technologies, and focusing on cardiac disease detection, this study offers valuable insights into the development of a robust and accurate machine learning system for disease prediction.

Keywords: Artificial Intelligence, Big Data, Cardiac Diseases, EMRBots, Data Preprocessing, Machine Learning Algorithms, Pandas, Scikit-Learn, NumPy, Accuracy, Precision, Recall, F1 score

Agradecimientos

En primer lugar, agradezco a la Universidad Politécnica de Madrid y a cada uno de los profesores que con paciencia y motivación, nos enseñaron todos los conocimientos para formarnos en el ámbito de la ingeniería biomédica y distintos valores y virtudes que nos serviran para toda nuestra vida.

Por otro lado, agradezco a mis amigos y compañeros, tanto de la universidad como de mi país natal Venezuela, que me han apoyado durante los momentos más duros y serán mis hermanos para toda la vida.

Finalmente, agradezco a una de las personas mas listas que he conocido en mi vida, mi tutor Oscar Araque por haber sido capaz de ver mi potencial de trabajo en el ámbito tecnológico y por su confianza y gran ayuda durante el desarrollo de este trabajo.

Sin todos ellos no habría sido posible desarrollar este trabajo y graduarme como Ingeniero Biomédico.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

The world is experiencing a dynamic era characterized by rapid advancements in technology, particularly in the fields of artificial intelligence (AI) and big data analytics [1]. These innovations have revolutionized various industries, and healthcare is no exception. With the availability of vast amounts of data and the computational power to process it, researchers and healthcare professionals now have unprecedented opportunities to improve disease prediction, diagnosis, and treatment.

The healthcare sector is facing numerous challenges, such as an aging population, an increase in chronic diseases, and the need for more personalized and efficient healthcare approaches. Traditional healthcare practices often rely on subjective assessments and limited data, leading to variations in diagnoses and treatment outcomes. However, the integration of AI and big data analytics has the potential to transform healthcare by providing data-driven insights, enhancing decision-making processes, and ultimately improving patient outcomes [2].

In this context, the proposed research on the design and development of a machine learning system for disease prediction holds great relevance. By harnessing the power of

machine learning algorithms and leveraging demographic and laboratory data, the aim is to develop a highly accurate model capable of predicting diseases and enabling personalized healthcare approaches. This aligns with the global movement towards precision medicine, which focuses on tailoring healthcare interventions to individual characteristics, thereby maximizing effectiveness and minimizing adverse effects.

Furthermore, the ongoing COVID-19 pandemic [3] has underscored the importance of early detection, accurate diagnosis, and efficient healthcare delivery. The availability of a robust machine learning system for disease prediction can significantly contribute to mitigating the impact of future pandemics or other public health emergencies.

Therefore, the proposed research is situated within a context of technological advancements, increasing healthcare challenges, and the urgent need for more accurate and personalized approaches. By capitalizing on machine learning and data analytics, this study seeks to make a small, but interesting contribution to the field of predictive medicine, offering a study insights and tools that can help disease prediction and healthcare delivery in the contemporary world.

1.2 Project goals

This work's original main goal was to create a prediction model that could accurately predict different diseases using demographic information and laboratory data. However, as the study went on, it became clear that concentrating on a particular ailment would be more doable and useful. In this instance, it was decided to focus on the likelihood of heart disease.

The study was able to look deeper into the precise traits, risk factors, and patterns linked with this particular illness by focusing on heart disease prediction. This method made it possible to analyze the available data more narrowly and create a model that was specifically designed to predict cardiac disease with better results.

Because of its huge impact on world health and the potential advantages of early identification and intervention, heart disease was chosen as the focus. This particular ailment was the focus of the research, which attempted to offer insightful information about the risk factors, signs, and patterns of heart disease. This knowledge might then be applied to enhance diagnosis, create individualized healthcare plans, and ultimately aid in lessening the impact of heart disease on people's lives and healthcare infrastructure. The objectives to be developed during the work are:

1. Obtain complete datasets containing real values on physiological, demographic, or analytical variables and perform in-depth analysis and preprocessing of the data.

2. Develop and evaluate multiple prediction models using different datasets, aiming to achieve an f1-macro accuracy above 80 percent, and gain an internal understanding of the model's learning process.

3. Gain insights into the physiological significance of the model's key characteristics, enhancing understanding of underlying patterns and relationships.

1.3 Structure of this document

In this section we provide a brief overview of the chapters included in this document. The structure is as follows:

Chapter 1 provides a concise yet comprehensive understanding of the work's objectives, the problems it aims to address, and the approach that will be taken to tackle these challenges. It serves as a roadmap for the reader, outlining the motivation behind the project, identifying the gaps in existing knowledge or practices, and setting expectations for the subsequent sections where the research methodology and findings will be presented.

Chapter 2 showcases the tools and platforms utilized in its development, emphasizing their key features and functionalities that enable effective problem-solving. It highlights the strategic selection of technologies tailored to address the identified issues.

Chapter 3 provides comprehensive insights into the research outcomes, presenting an elaborate exposition of the project's global architecture and intricate components. It delves into the intricate details of the achieved results, showcasing their significance and implications in the realm of scientific inquiry.

Chapter 4 guides the reader into an immersive exploration of the application, captivating their attention through vivid textual and visual depictions of the user's interactions within the dynamic graphical interfaces meticulously crafted using Pyplot. It presents diverse use cases, meticulously portraying the myriad actions performed by the user, evoking a sense of engagement and fostering a deeper understanding of the application's functionality.

Chapter 5 concludes the work with the conclusions extracted from the project, analyzing the overall results with the achieved goals and discussing future improvements that the application could have in the future.

CHAPTER 2

Enabling Technologies

In this chapter, we provide an overview of the enabling technologies that play a crucial role in the development and implementation of our project. These technologies serve as the foundation for the successful execution of our objectives and contribute to the overall effectiveness and efficiency of our solution.

2.1 Python Ecosystem

The discipline of biomedical engineering has seen a revolution because to the flexible and popular computer language Python [4]. Python's popularity increased when it was created in the late 1980s thanks to its ease of use, readability, and wide range of libraries and frameworks.

Several important elements have contributed to Python's influence in the field of biomedical engineering [5]. The first benefit is that academics and engineers may create sophisticated algorithms and analyze enormous datasets more easily thanks to its intuitive and expressive vocabulary. The biomedical engineering community benefits from effective collaboration and code exchange thanks to Python's readability.

Another key factor in Python's success in biomedical engineering is its vast ecosystem of libraries and frameworks. For numerical computing, data analysis, and manipulation, sophisticated capabilities are available from libraries like NumPy, SciPy, and Pandas. Engineers and scientists may easily complete complex data processing, statistical analysis, and modeling jobs thanks to these packages.

Python's influence on the biomedical engineering sector has also been influenced by how well-liked and widely used it is in the scientific community. Advancements in fields like medical imaging, signal processing, and pattern recognition have been sped up by the availability of specialized libraries and frameworks for tasks like machine learning (e.g., Scikit-Learn and TensorFlow) and image processing (e.g., OpenCV) [6].

2.1.1 Jupyter Notebook

Jupyter Notebook [7] enables the programmer create and share documents with live code, graphics, and explanations. For interactive work and data analysis, it is frequently utilized by researchers, educators, and data scientists.

Furthermore, it works well with well-known scientific computing and data analysis tools like NumPy, Pandas, Matplotlib, and SciPy. These libraries offer access to a broad ecosystem of tools and features for data processing, analysis, and visualization and are simple to import and use within notebook cells.

2.1.2 Scikit-Learn

Scikit-Learn [8] is an open-source machine learning library for Python that goes by the name of sklearn. By providing healthcare practitioners and academics with potent tools for predictive modeling, data analysis, and pattern identification, it has significantly contributed to the modernizing of the medical field.

Numerous machine learning algorithms, including supervised and unsupervised learning techniques, as well as tools for data preprocessing, model selection, and evaluation, are available through Scikit-Learn. These algorithms and technologies give medical practitioners the ability to glean insightful information from medical datasets, form precise forecasts, and create cutting-edge decision support systems.

Additionally, Scikit-Learn has made it easier to combine machine learning with clinical data and electronic health records (EHR) [9]. Healthcare experts may find hidden patterns in massive amounts of patient data, identify risk factors, and create individualized treatment

programs by utilizing unsupervised learning methods like clustering and dimensionality reduction.

2.1.3 Pandas

The discipline of biomedical engineering has seen a revolution thanks to Pandas, an open-source package for Python [10] that allows for data processing and analysis.

It has become a useful tool for biomedical researchers and engineers because to the emergence of data structures like Series and DataFrame, created for the effective processing of structured data. Working with real-world biomedical datasets has necessitated the use of Pandas because of its capacity to manage missing data and carry out data cleaning and preprocessing operations.

2.1.4 NumPy

The Python ecosystem's core library, NumPy [11] —short for Numerical Python— has had a substantial impact on a number of fields, including biomedical engineering. Along with a range of mathematical methods to manipulate these arrays, it offers a strong and effective method for handling sizable, multidimensional arrays and matrices. The field of numerical computation and data analysis has been completely transformed by this capability.

Due to its smooth interface with other Python scientific and data analysis libraries like Pandas, SciPy, and Matplotlib, NumPy has a wide range of applications. This allows for a whole data analysis workflow. NumPy serves as a basis for enabling engineers and researchers in the biomedical area to carry out sophisticated mathematical operations, effectively work with arrays, and use a variety of algorithms and functions to tackle challenging problems.

2.1.5 PyPlot

A robust Python plotting toolkit called PyPlot [12] offers a variety of tools and features for building visually beautiful and educational plots. It offers a streamlined interface for creating numerous plot types, including line plots, scatter plots, bar plots, histograms, and more. It is developed on top of the Matplotlib package.

With PyPlot, users can more easily understand and interpret the underlying patterns and trends in their data since complicated data and computational concepts can be visually

communicated. PyPlot makes data exploration and analysis easier by charting data points, trends, and relationships, assisting in the discovery of key characteristics and insights.

2.2 Background

The idea behind the thesis, arises from the expanding demand for precise and effective patient sickness prediction systems. Early and accurate diagnosis is essential for enhancing patient outcomes, treatment planning, and resource allocation in the healthcare sector. Traditional diagnostic techniques may be prone to mistakes or delays since they frequently rely on subjective judgements.

Our effort seeks to address these issues and develop predictive medicine by utilizing machine learning techniques and studying demographic and laboratory data. The objective is to create a system that can accurately forecast patient illnesses based on a wide range of characteristics, such as demographic data and laboratory test results.

The development of machine learning techniques, the accessibility of big and varied datasets, and the growing emphasis on data-driven methods in healthcare provide the foundation for this study. These elements working together have created new opportunities for building precise and trustworthy forecasting models.

Our initiative seeks to contribute to the developing field of predictive medicine, where machine learning can play a crucial part in revolutionizing healthcare practices, by utilizing these technology breakthroughs and examining the vast demographic and laboratory data readily available.

The use of machine learning algorithms to forecast patient ailments has the potential to be advantageous in the future. It can aid medical professionals in spotting high-risk patients who might need immediate care or intervention, resulting in prompt medical interventions and possibly saving lives.

Machine Learning Applied to Healthcare: The future of Medicine

In my opinion, the future of healthcare will be shaped by machine learning as it is applied to the field of medicine. Large and diverse datasets combined with effective machine learning algorithms have the potential to reveal important insights, advance diagnosis, individualize therapies, and improve patient outcomes.

First, a significant resource for analysis is made available by the abundance of healthcare data, which includes electronic health records, medical imaging, genomics, wearable devices,

and more. These databases contain important data about the patient's demographics, medical history, disease patterns, response to treatment, and results. Machine learning algorithms can use this data to find trends, correlations, and patterns that may not be visible to human therapists.

Additionally, machine learning can be extremely important in precision medicine by allowing doctors to customize each patient's treatment plan based on their particular traits. Machine learning algorithms can determine the most effective treatment plans, anticipate drug responses, and reduce bad reactions by taking into account a patient's genetic profile, lifestyle characteristics, and medical history.

Healthcare systems that include machine learning could result in better patient outcomes, greater efficiency, and lower costs. It can facilitate early intervention, stop pointless procedures, and expedite administrative processes in the medical field. Machine learning may also help with resource allocation, forecasting resource needs for healthcare, and improving healthcare delivery.

CHAPTER 3

Resources

3.1 Datasets

A dataset [13] is a grouping of organized and structured or unstructured data used for analysis, research, or other purposes. It can be viewed as a "container" for associated data or observations. In the context of data science and machine learning, a dataset typically refers to a structured collection of data that is used for training, testing, and evaluating models.

During the development of the work we will be working with three different types of datasets, the first, called EMRBots, is an artificial generation (with medical sense) to create medical records with laboratory tests and demographic data of the patients. As previously indicated, throughout the work it was decided to specify a disease (heart disease), so we also have two datasets focused on heart disease, such as Cardiotocography (fetal cardiotocograms) and Heart Disease (diagnosis of heart disease).

Initial Overview

As has been commented throughout the development of the work, we will be working with three datasets in particular, EMRBots, Cardiotocography and Heart Disease. Next,

we will not only do an extensive explanation of the datasets and their composition, but we will also start the analysis with a comparison between them to understand the main differences.

	EMRBots	**Cardiotocography**	**Heart Disease**
Topic	Demographic and Laboratory	Fetal Cardiotocograms	Cardiac Diseases
Size	10,000	2,126	303
Initial Variables	18	21	14
Data Type	Mixed	Numerical	Numerical
Correlation	Low	Medium	Low
Class Balance	Imbalanced	Imbalanced	Balanced

3.1.1 EMRBots

Artificially created electronic medical records, or EMRBots [14], are used experimentally. The purpose of EMRBots is to make it possible for non-commercial organizations to train statistical and machine-learning algorithms using the fictitious patient repository. As long as the repositories are are not utilized to produce commercial software goods, commercial enterprises are likewise permitted to use the repositories for any purpose.

Sensitive personal data is kept in EMRs. For instance, they could offer information about mental illnesses or specifics about contagious diseases like the human immunodeficiency virus (HIV). They might also include other private information, like specifics about fertility treatments' medical conditions. Only a select few people are permitted to access and analyze EMR databases due to confidentiality regulations that apply to EMRs. Employees at organizations without access to EMR systems are unable to gain practical experience with this important tool.

An important study and to take into account about EHRBots is Health-ATM [15] that is an academic paper that was presented at the SIAM International Conference on Data Mining. The project's goal was to create the Health-ATM deep learning architecture, which can precisely anticipate patient health risks and properly represent complex patient health records.

The Health-ATM architecture, which stands for Health-Adaptive Tensor Machines, was suggested by the researchers for this project. For the purpose of capturing the underlying relationships and patterns in the patient health records, the architecture makes use of deep learning techniques, notably deep tensor factorization. The entire deep learning process comes from the dataset which we will be working on throughout the project.

This architecture can improve patient care and results and has substantial implications for the healthcare industry. Healthcare practitioners can make more educated decisions, personalize treatments, and proactively intervene to prevent or lessen health hazards by accurately representing and analyzing patient health records.

The dataset contains a set of 10,000 anonymous patients, which is made up of the following files:

3.1.1.1 PatientCorePopulatedTable

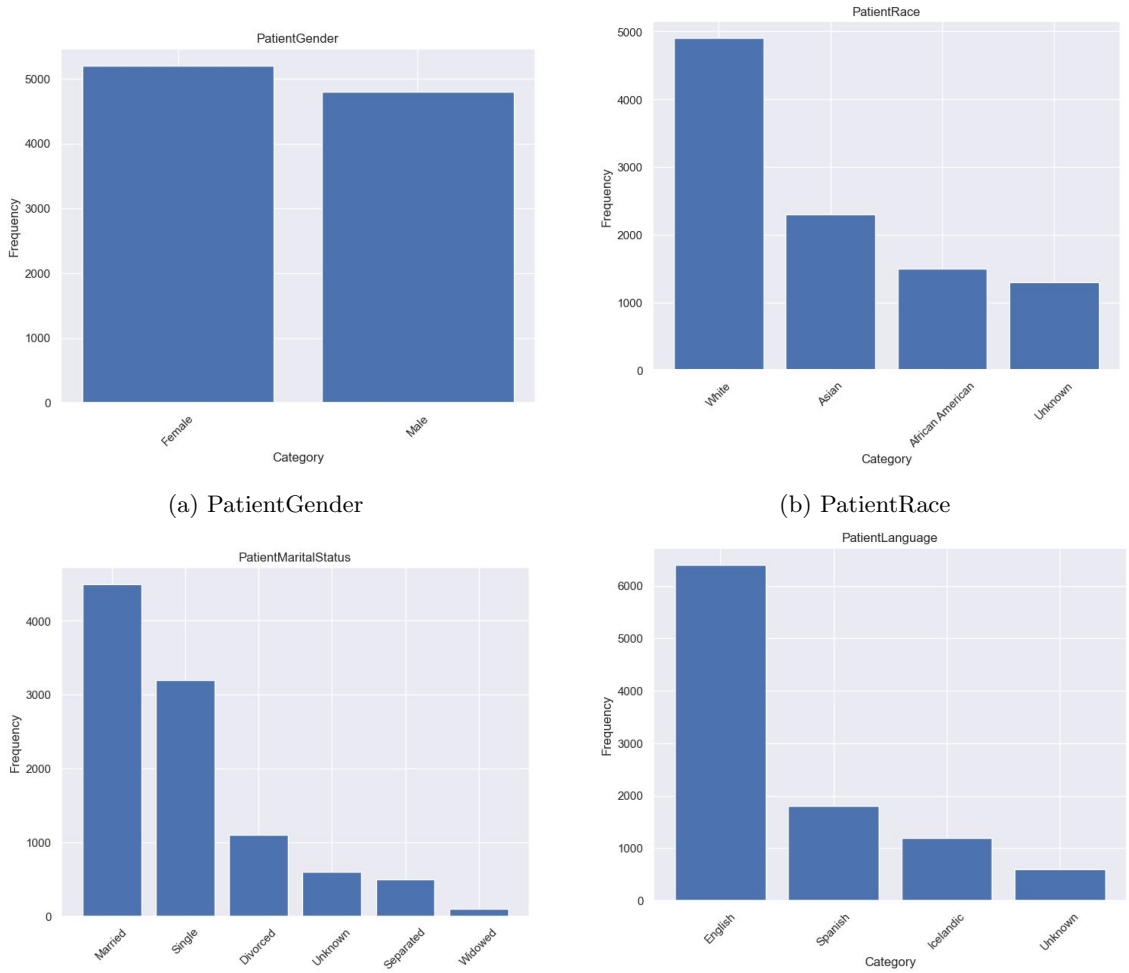
PatientID	a unique ID representing a patient
PatientGender	Male/Female
PatientDateOfBirth	Date Of Birth
PatientRace	African American, Asian, White
PatientMaritalStatus	Single, Married, Divorced, Separated, Widowed
PatientLanguage	English, Icelandic, Spanish
PatientPopulation%BelowPoverty	given in %

The "PatientCorePopulatedTable" dataset contains essential information about patients. Each patient is identified by a unique ID ([PatientID]). The dataset includes columns such as [PatientGender] indicating the gender of the patient (Male/Female), [PatientDateOfBirth] representing the date of birth, [PatientRace] specifying the racial background (African American, Asian, White), [PatientMaritalStatus] indicating the marital status (Single, Married, Divorced, Separated, Widowed), [PatientLanguage] indicating the language spoken by the patient (English, Icelandic, Spanish), and [PatientPopulationPercentageBelowPoverty] representing the percentage of the patient population living below the poverty line.

This dataset provides valuable demographic and socioeconomic information about the

patients, enabling analysis and exploration of various factors related to healthcare and patient outcomes. Below is an exhaustive analysis of the distribution of the variables:

As can be seen below, Figure 3.1 shows the distribution of the categorical variables of the PatientCorePopulatedTable table. It is observed that the sex is very well equitable, with slightly more women than men, but both PatientRace, PatientMaritalStatus and PatientLanguage are noticeably skewed to the left.



(a) PatientGender

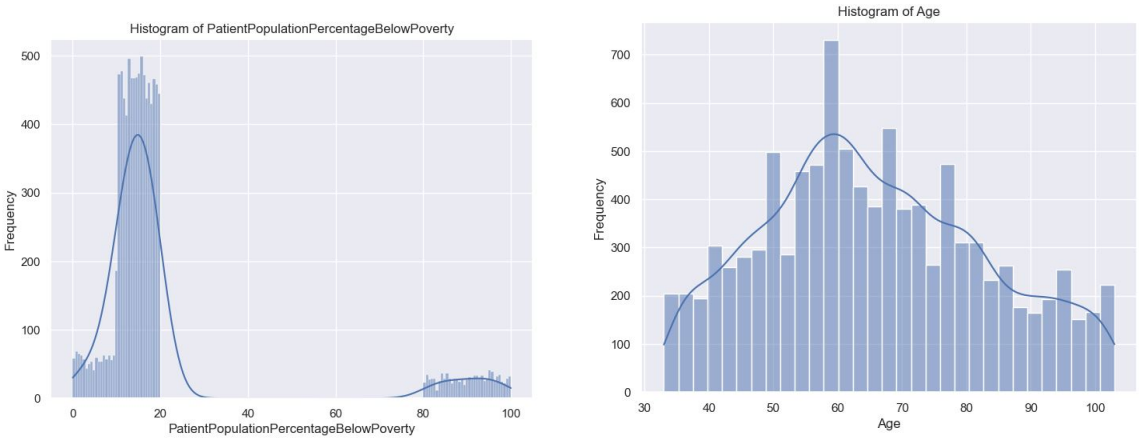
(b) PatientRace

(c) PatientMaritalStatus

(d) PatientLanguage

Figure 3.1: Categorical Variables of PatientCorePopulatedTable

On the other hand, Figure 3.2 shows the distribution of the numeric variables of the PatientCorePopulatedTable table. It is observed that the variable that proposes the percentage below poverty is quite biased to the left as shown by the moving average and the age is quite equitable, being 60 years the most repeated age.



(a) PatientPopulationPercentageBelowPoverty

(b) Age (PatientDateOfBirth)

Figure 3.2: Numerical Variables of PatientCorePopulatedTable

3.1.1.2 AdmissionsDiagnosesCorePopulatedTable

PatientID	a unique ID representing a patient
AdmissionID	an admission ID for the patient
PatientDateOfBirth	Date Of Birth
PrimaryDiagnosisCode	ICD10 code for admission's primary diagnosis
PrimaryDiagnosisDescription	admission's primary diagnosis description

In this dataset, the variable that we are going to predict called Primary Diagnosis Code is made up of a specific code. ICD-10 [16] is the tenth update to the World Health Organization's (WHO) International Statistical Classification of Diseases and Related Health Problems (ICD), a list of medical classifications. It includes codes for physical characteristics, unusual observations, social contexts, and external causes of harm or illness. In this last table of values there is no correct distribution which can be visualized with a histogram, since there is no clear distribution

3.1.1.3 LabsCorePopulatedTable

PatientID	a unique ID representing a patient.
AdmissionID	an admission ID for the patient.
LabName	lab's name, including the names of the values sought a standard basic blood test for any human.
LabValue	lab's value
LabUnits	lab's units
LabDateTime	date

The "LabsCorePopulatedTable" dataset consists of patient-related information and laboratory test results. Each patient is assigned a unique ID ([PatientID]) and is associated with an admission ID ([AdmissionID]). The dataset includes columns such as [LabName] for the name or identifier of the laboratory test, [LabValue] for the numerical test result, [LabUnits] for the units of measurement, and [LabDateTime] for the date and time of the test. This dataset provides valuable information for analyzing and studying the outcomes of various laboratory tests conducted on patients.

3.1.2 Cardiotocography

Moving to the next dataset [17], contains 2126 fetal cardiotocograms belonging to different classes. The dataset consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on cardiotocograms classified by expert obstetricians.

An important study and to take into account about this specific Cardiotocography dataset is a research study with the working title "Fetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements" [18] in the IEEE Xplore Digital Library. The project's goal was to create a decision tree classifier-based system for tracking fetuses' health conditions using data from cardiotocography.

Using CTG readings, the researchers' goal in this effort was to create a classification algorithm based on decision trees that could precisely predict the fetal health state. They made use of a set of CTG measures obtained from expectant mothers, which included

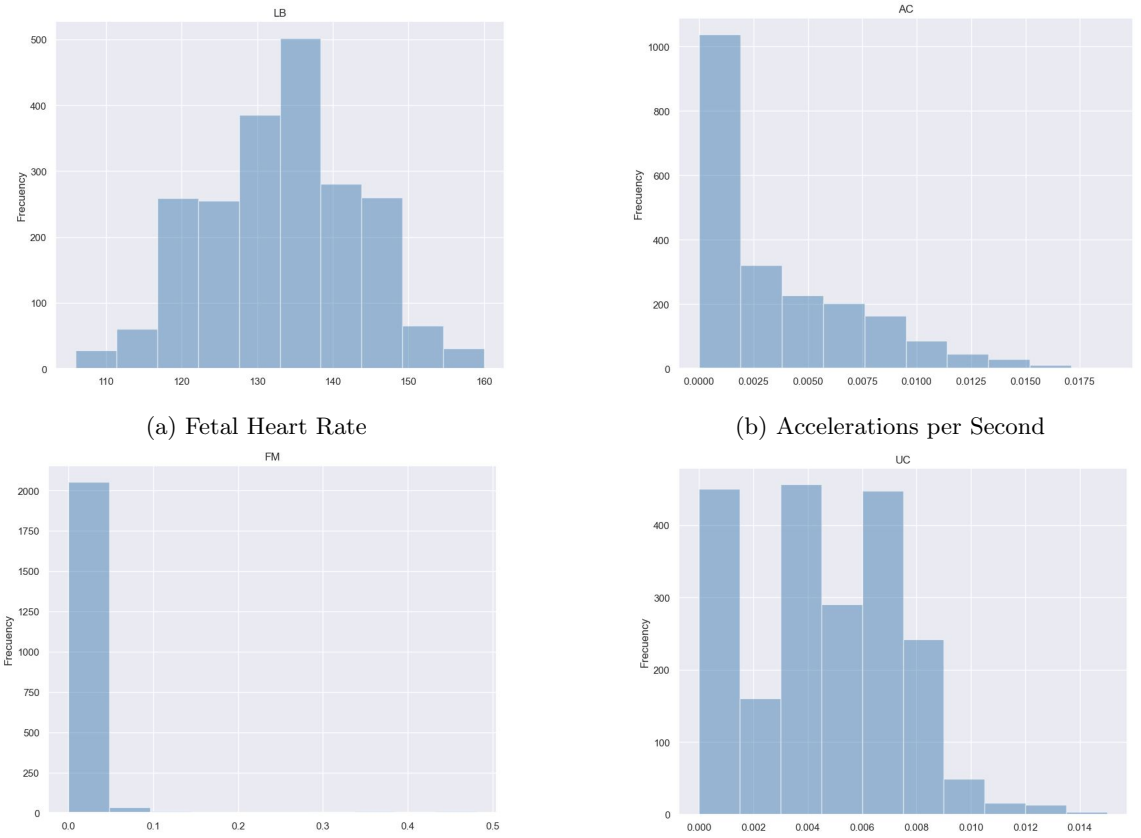
characteristics including the fetal heart rate at rest, accelerations, and decelerations, as well as uterine contractions.

This data set consists of 21 attributes which include:

LB	FHR baseline (beats per minute)
AC	accelerations per second
FM	fetal movements per second
UC	uterine contractions per second
DL	light decelerations per second
DS	severe decelerations per second
DP	prolonged decelerations per second
ASTV	percentage of time with abnormal short term variability
MSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
MLTV	mean value of long term variability
Width	width of FHR histogram
Min	minimum of FHR histogram
Max	Maximum of FHR histogram
Nmax	histogram peaks
Nzeros	histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance

Tendency	histogram tendency
CLASS	FHR pattern class code (1 to 10)
NSP	Normal=1; Suspect=2; Pathologic=3

As can be seen below, in the Figure 3.3 in subfigure 3.3a, it can be seen that there is a uniform distribution of the fetal heartbeat, with the largest number of values at 135bpm. In subfigure 3.3b, it can be seen that the accelerations of the fetus per second are biased to the left, as well as in subfigure 3.3c with the movements of the fetus per second, being very small values since the fetuses do not move as much per second. Finally, in subfigure 3.3d, uterine contractions per seconds, it can be seen that it is slightly biased to the left, with very small values, since uterine contractions happen less often.



(a) Fetal Heart Rate

(b) Accelerations per Second

(c) Fetal Movements per second

(d) Uterine Contractions per second

Figure 3.3: Distribution of Cardiotocography relevant variables

3.1.3 Heart Disease

This data set [19] contains 4 databases related to the diagnosis of cardiac diseases. All the attributes have numerical values, fully anonymized, but they are real data, from real life, from real patients.

A surprising research article with the working title "A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset" [20] was released in the Springer publication. The project's goal was to create a hybrid machine learning method for diagnosing coronary disorders utilizing a feature selection technique used on a dataset of heart disease.

The results of this study have implications for cardiovascular health since they show how machine learning approaches can help with coronary disease detection and diagnosis. Utilizing the heart disease dataset to promote early identification and intervention in patients with coronary illnesses is made possible by the hybrid approach.

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. Notably, the **Cleveland database** is the only one that has been used by machine learning researchers to date. The "goal" field refers to the presence of cardiac disease in the patient. It has an integer value from 0 (no presence) to 4.

As a result, it was decided to work with the "Cleveland" dataset which include 303 instances. since it is the dataset with the most data and the one that the creators of the dataset recommend to use. The RAW dataset contains 76 attributes which will not all be specified because it is quite long, only the 14 attributes will be specified which we are going to work, since the others are not relevant. The attribute information is:

Age	pacient's age
Sex	[1 = male; 0 = female]
cp	"Chest Pain Type" - [Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic]
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholesterol in mg/dl

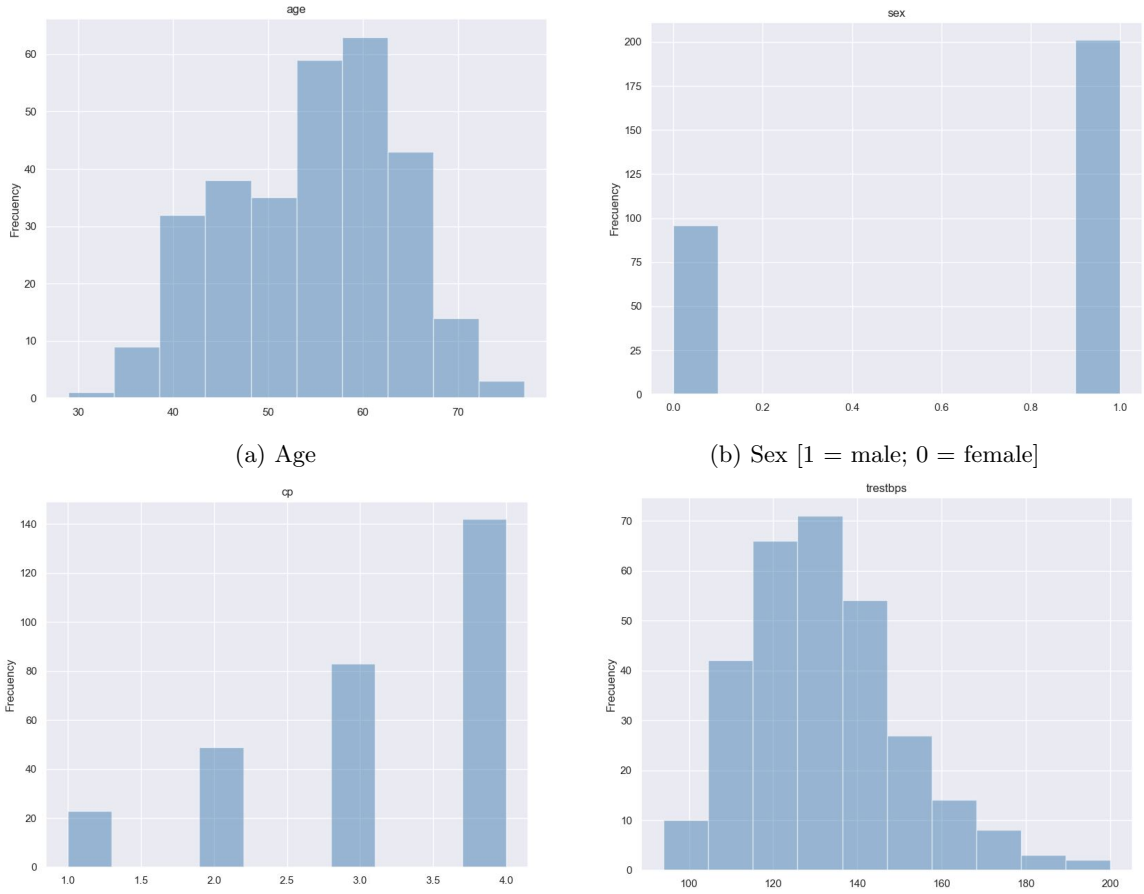
fbs	(fasting blood sugar over 120 mg/dl) [1 = true; 0 = false]
restecg	resting electrocardiographic results - [Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of ¿ 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria]
thalach	maximum heart rate achieved
exang	exercise induced angina [1 = yes; 0 = no]
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment [Value 1: upsloping, Value 2: flat, Value 3: downsloping]
ca	number of major vessels (0-3) colored by flourosopy
thal	[3 = normal, 6 = fixed defect, 7 = reversable defect]
num	diagnosis of heart disease (angiographic disease status) - [Value 0: negative heart disease b) Value 1: positive heart disease]

As we can see, the dataset is 100% focused on detecting heart disease, but in this case, mainly in adults, the last variable being the variable to predict, seeking to find out if the human has heart disease or not. Below is an exhaustive analysis of the distribution of the variables:

As can be seen below, Figure 3.4 shows the distribution of important variables in the dataset. In subfigure 3.4a it can be seen that age has a uniform distribution, slightly skewed to the right with the age of majority located between 50 and 60. On the other hand, in subfigure 3.4b it can be seen that there are twice as many men in the dataset what women

Subfigure 3.4c shows the distribution of the chest pain type, with value 4 being the most

predominant, and subfigure 3.4d shows the resting blood pressure of the patients, where the highest values are in normal values.



(a) Age

(b) Sex [1 = male; 0 = female]

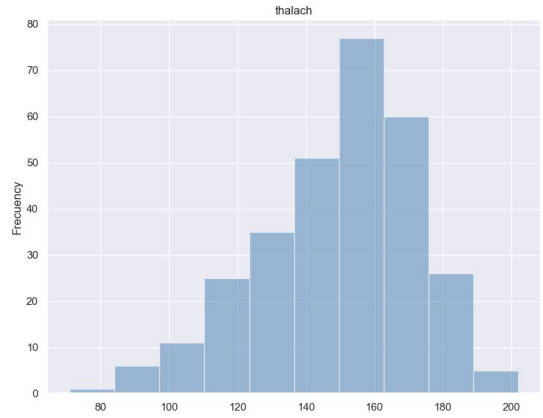
(c) Chest Pain Type

(d) Resting Blood Pressure

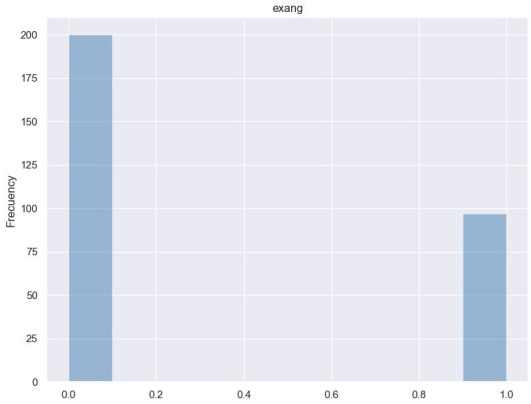
Figure 3.4: Distribution of Heart Disease variables 1

Finally, other variables to take into account included in the Figure 3.5 such as subfigure 3.5a, shows the maximum heart rate achieved by the patients, subfigure 3.5b if it is the case of angina pectoris generated by exercise and subfigure 3.5c which marks the number of blood vessels (important) colored by fluoroscopy. These three variables play a vital role in the medical detection of heart disease.

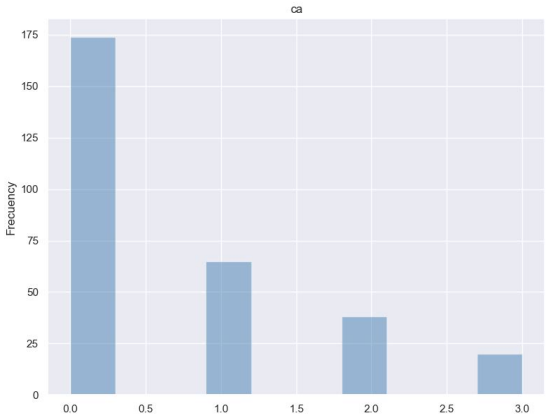
On the other hand, in this case the variable to be predicted (subfigure 3.5d) has two outcomes, 0 is equal to no heart disease and 1 affirms heart disease, in this case, it is observed that the distribution of the variables is quite equitable.



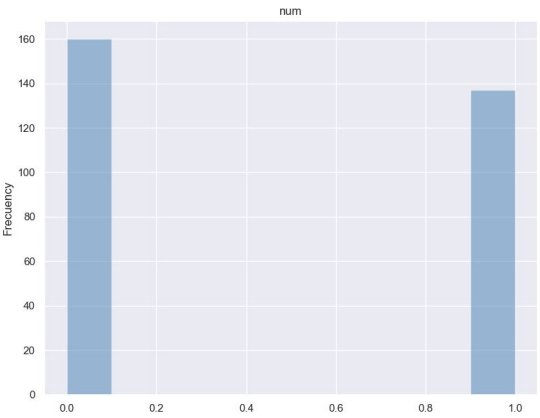
(a) Maximum Heart Rate Achieved



(b) Exercise Induced Angina [1 = yes; 0 = no]



(c) Number of Major Vessels colored by flourosopy



(d) Variable to predict

Figure 3.5: Distribution of Heart Disease variables 2

CHAPTER 4

Evaluation

4.1 Methodology

A crucial phase in data analysis, is data processing which is essential for obtaining insightful information and knowledge. A systematic strategy to handling and processing datasets is provided by dataset treatment methodology, ensuring its quality, integrity, and relevance for later analysis [21]. This methodology includes a number of processes, such as feature selection, data cleansing, preprocessing, integration, and integration.

Data cleaning entails locating and fixing mistakes, discrepancies, and missing values in the dataset [21]. Preprocessing concentrates on eliminating noise, standardizing and normalizing the data, and addressing outliers. Integration is the process of fusing various datasets from many sources into a single format for analysis. Data must be transformed into a suitable representation using techniques like scaling, encoding, or dimensionality reduction. Last but not least, feature selection seeks to pinpoint the most pertinent and instructive features that support the analysis's goals.

In the case of the three datasets that we have mentioned throughout this work, we have applied preprocessing techniques to all of them to obtain the most optimal information when training the corresponding machine learning model.

For all datasets, the *Decision Tree Classifier*, *Random Forest Classifier* and *Support Vector Machine (SVM)* algorithms have been used [22].

Decision Tree Classifier

The decision tree approach for classification tasks is implemented by the Decision Tree Classifier class in the Python Scikit-learn module [23]. It is a component of the larger collection of machine learning algorithms offered by Scikit-learn, which is frequently used for modeling and data analysis.

The conventional decision tree approach is used by Scikit-learn's Decision Tree Classifier to recursively split the data based on feature values and produce a tree-like structure. Based on certain criteria, such as maximizing information gain or decreasing impurity measurements like the Gini index or entropy, the algorithm chooses the appropriate feature to split the data at each stage. Until a stopping requirement is met, such as a maximum tree depth or a minimum amount of samples in each leaf node, this splitting process is carried out.

Scikit-learn's Decision Tree Classifier offers various hyperparameters [24] that can be tuned to control the tree's complexity, such as the maximum depth of the tree, the minimum number of samples required to split an internal node, or the minimum number of samples required to be at a leaf node. Additionally, it provides features like handling categorical features, handling missing values, and pruning the tree to prevent overfitting.

With a user-friendly interface, effective implementation, and the capacity to handle both numerical and categorical data, Scikit-learn's Decision Tree Classifier is an all-around effective and adaptable tool for classification jobs. Making predictions and learning from data is a common practice in many industries, including healthcare, finance, and natural language processing.

Random Forest Classifier

The Random Forest Classifier in Scikit-learn [25] generates a group of decision trees by randomly choosing portions of the training data and features. On these subsets, each decision tree is trained independently, and during prediction, the combined predictions of the multiple trees decide the final result. For classification tasks, the class with the most votes is chosen. For regression tasks, this aggregation can be carried out by averaging the projected probabilities.

A range of hyperparameters, including the number of ensemble trees, the maximum depth of each tree, and the amount of features to take into account when looking for the

optimum split, are available for the Random Forest Classifier in Scikit-learn [26]. The Random Forest Classifier can lessen overfitting and increase the model's capacity for generalization by mixing many decision trees.

Overall, Scikit-learn's Random Forest Classifier is a strong and adaptable tool for classification tasks, utilizing the combined knowledge of several decision trees to provide more accurate predictions and successfully manage large datasets.

Support Vector Machine (SVM)

The Python Scikit-learn library uses the Support Vector Machine (SVM) machine learning technique [27]. It is a potent supervised learning technique used for both regression and classification tasks. In complex issues with non-linear decision boundaries, SVMs are especially useful.

The SVM method in Scikit-learn looks for an ideal hyperplane that maximally divides various classes or groupings of data points. The data are divided into multiple groups by this hyperplane, which serves as a decision boundary [28]. SVMs are renowned for their proficiency with high-dimensional feature spaces and for delivering good results even when the number of dimensions is greater than the number of samples.

SVMs also offer a number of hyperparameters that may be changed to enhance the model's performance, such as the kernel-specific parameters and the regularization parameter C, which regulates the trade-off between maximizing the margin and reducing the classification error.

Finally, Support Vector Machine provides a flexible and effective technique for handling classification and regression issues. It is a well-liked option in a variety of industries, including healthcare, banking, and natural language processing, thanks to its capacity to manage high-dimensional data, discover optimal decision limits, and support multiple kernel functions.

Classification Report

Composed by precision, recall, and F1-score are common performance metrics used in classification tasks. Precision represents the proportion of correctly predicted positive instances out of all instances predicted as positive, while recall measures the proportion of correctly predicted positive instances out of all actual positive instances. F1-score combines precision and recall into a single metric that provides a balanced measure of a model's performance, considering both false positives and false negatives.

Comparative table between Decision Tree Classifier, Random Forest Classifier, and Support Vector Machine (SVM):

Algorithm	**Pros**	**Cons**	**Best Use Case**
Decision Tree Classifier	- Interpretable and easy to understand - Handles both numerical and categorical features - No need for feature scaling - Works well with both binary and multi-class problems	- Prone to overfitting and high variance - Limited ability to capture complex relationships - Lack of robustness to small perturbations in the data	- Small to medium-sized datasets - Problems with non-linear decision boundaries - Exploratory data analysis, feature selection
Random Forest Classifier	- Reduces overfitting through ensemble of decision trees - Handles large datasets with high accuracy - Robust to noisy data and outliers - No feature scaling required	- Increased computational complexity and training time - Less interpretable compared to a single decision tree - Requires tuning of hyperparameters	- Classification tasks with high-dimensional feature space - Problems with non-linear decision boundaries - Handling imbalanced datasets, feature importance analysis

Algorithm	Pros	Cons	Best Use Case
Support Vector Machine (SVM)	- Effective in high-dimensional feature spaces	- Computationally intensive and memory-consuming	- Classification tasks with small to medium-sized datasets
	- Can handle non-linear decision boundaries	- Sensitive to choice of kernel function and parameters	- Problems with non-linear decision boundaries
	- Robust to outliers and noise	- Requires feature scaling	- Text classification, image recognition, bioinformatics
	- Can handle both binary and multi-class problems		

EMRBots Preprocessing

Undoubtedly the most complex preprocessing of the three datasets, since we were presented with the data divided into three different tables and many of them were not in numbers but in categorical variables of the"object" or "string" type.

First, we focused on transforming the "PatientDateOfBirth" variable to a new variable called "Age" so that we can better use it in preprocessing development. On the other hand, we eliminate certain irrelevant variables that we are not going to use such as LabDateTime, PatientDateOfBirth, LabUnits, PrimaryDiagnosisDescription and PatientID.

Second, we must simplify the PrimaryDiagnosisCode (Figure 4.1) so that the variables to predict are smaller. So we grouped certain variables that represented a common disease and eliminated others that did not appear frequently in the data set to make it as equitable as possible. Leaving only the letters C (Malignant neoplasms), L (Diseases of the skin and subcutaneous tissue), F (Mental and behavioral disorders), E (Endocrine, nutritional and metabolic diseases), and O (Pregnancy, childbirth and the puerperium) [29].

Afterwards, having 3 different datasets, we have to find a way to join them, so that we can work with only one. To do this, we use the Pandas merge [30] function, which allows us to join two dataframes, explaining which dataset we are going to rely on (right or left) and which variable we are going to focus on so that there is an order in the merge. It

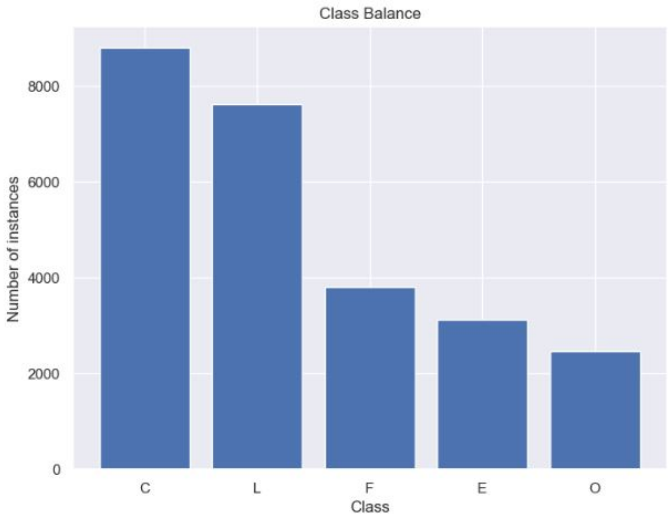


Figure 4.1: PrimaryDiagnosisCode (Simplified)

should be noted that it was decided to use merge over join [31] since the three tables had a different number of rows. First we do the merge of the PatientCorePopulatedTable with AdmissionsDiagnosesCorePopulatedTable.

When doing the previously mentioned merge, it generates a dataset that we call "df_extra", with which we are going to do a second merge with the dataset that is given to us in LabsCorePopulatedTable (df_diag). As in the previous merge, we leaned over the dataset on the right and specified to merge on PatientID and AdmissionID, in such a way that a single sorted table named df_final remains.

Then it is verified if there is a null variable with the "*isnull().any()*" function and in this case a drop of all the rows that have at least one null variable is performed to avoid problems, finally, we reset the index to format the new dataframe that we have generated from the old dataset.

The preprocessing continues, since the table must be pivoted [32], choosing as index: PatientID and AdmissionID, columns: LabName and as values: LabValues. Choosing the aggregate function of "mean" as an option and enabling the removal of any variable that does not exist. This step must be done, since each value of the laboratory tests must be a column in the dataset with its associated value when training the machine learning model.

As can be seen in Figure 4.2, almost all laboratory tests have a very uniform distribution, which facilitates learning about them.

Finally, a final merge is made of this pivoted dataset (df_pivot) and the dataset called

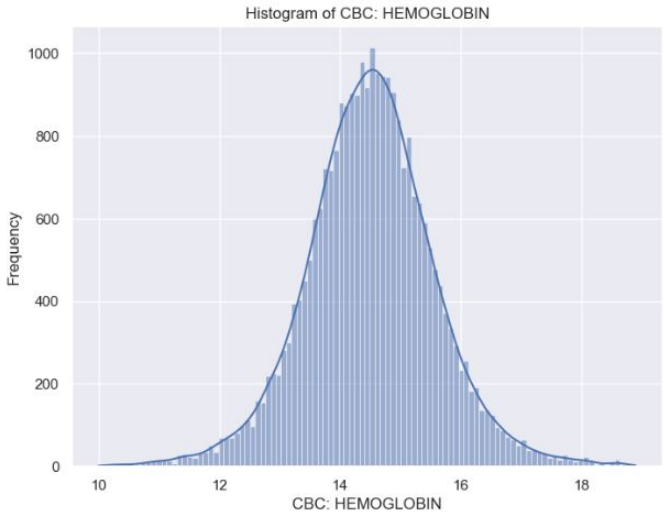


Figure 4.2: First example of uniform distribution

df_extra, always based on the ID variables (PatientID and AdmissionID) to obtain the last clean dataset.

Once the three datasets have been logically joined. We must encode the categorical variables, for this we use a preprocessing method called OneHotEncoder [33], which is used when encountering implicit ordering of variables to transform categorical data into a binary representation that preserves the distinction between different categories. For each distinct category, it constructs a new binary column and assigns a value of 1 if the data item fits into that category and a value of 0 otherwise. By transforming categorical data, machine learning algorithms may handle categorical data more effectively and avoid giving categories arbitrary numerical values.

The preprocessing method is applied to all the categorical variables of the dataset in order to apply the different machine learning algorithms. It should be noted that this preprocessing method generates a dataframe with the encoded values that we must concatenate with the pandas concat [34] function to the initial dataframe and eliminate the non-encoded column, as can be seen in the example:

Then, we must normalize the numerical values to improve the effectiveness of the prediction model, for this we use machine learning preprocessing methods like StandardScaler [35] that are used to uniformly scale numerical features in a dataset. The data are transformed into a distribution with a mean of zero and a variance of one by subtracting the mean from the data and dividing by each feature's standard deviation.

Because many machine learning algorithms are sensitive to the scale of the input vari-

ables, this technique makes sure that all features have a similar scale. By removing any potential bias or dominance that can be introduced by variables with various scales, StandardScaler helps to standardize the data. By using StandardScaler, the features are made more comparable so that the algorithm can anticipate outcomes with greater accuracy and dependability.

Once the preprocessing is complete, we need to train the model and predict. To do this, we divide the dataset into X, which contains all the dataset information except the variable to be predicted, and Y, which is, in effect, the variable to be predicted. Subsequently, it is divided into a test set and a training set with a ratio of 80% training, 20% test.

Cardiotocography Preprocessing

In the Cardiotocography dataset, it was fortunate to find that the data was already well optimized and required no preprocessing techniques, but a normalization of the data with Standard Scaler [35] has been carried out to further optimize the results. This dataset was carefully curated and prepared, ensuring that the variables and features were in a suitable format for analysis.

As a result, it was not necessary to perform any additional steps such as handling missing values or encoding categorical variables. The dataset was ready for direct use in our analysis, saving us valuable time and effort in the preprocessing phase.

The variable to predict in this dataset, NSP, contained 3 values whose meaning was Normal=1; suspect=2; Pathologic=3. As seen in Figure 4.3, the data was noticeably skewed to the left, with the majority of patients being in the "Normal" status dataset.

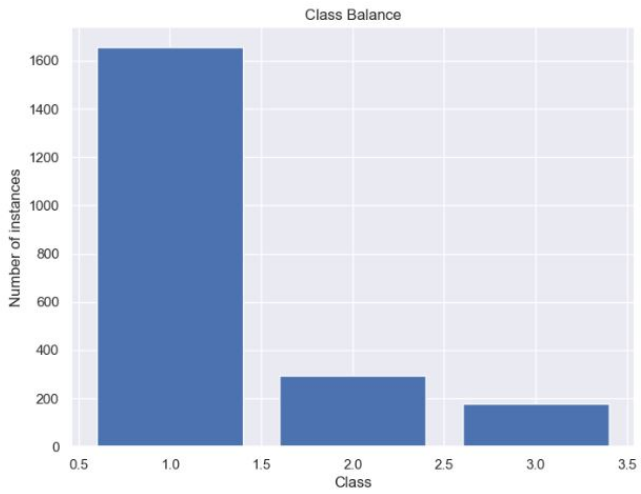


Figure 4.3: NSP Variable (to predict) Distribution

By having fewer columns, a study of the correlation matrix [36] was carried out, can be seen in the Figure 4.4 that is a table that shows the correlation rates between the elements of a dataset. It offers important details regarding the direction and strength of the linear relationship between two sets of variables. A value close to 1 suggests a strong positive correlation, a value close to -1 shows a strong negative correlation, and a value close to 0 indicates a weak or no association. The correlation coefficient has a range of -1 to 1.

In the case of the dataset we are analyzing, the correlation matrix does not exhibit a favorable pattern. This suggests that there is a lack of strong linear relationships between the variables.

A poor correlation matrix could indicate that the variables are not directly related or that other non-linear relationships may exist. It may pose challenges when applying certain statistical techniques or machine learning algorithms that assume linear relationships.

In figure 4.4, it can be seen that although most of the variables do not have a clear correlation, the variables “Width” and “Min” have a very strong inverse correlation with a value of -0.9. On the other hand, the variables “Mode”, “Mean” and “Median” all three have a very high correlation between them and also share an important correlation with the variable ”LB”.

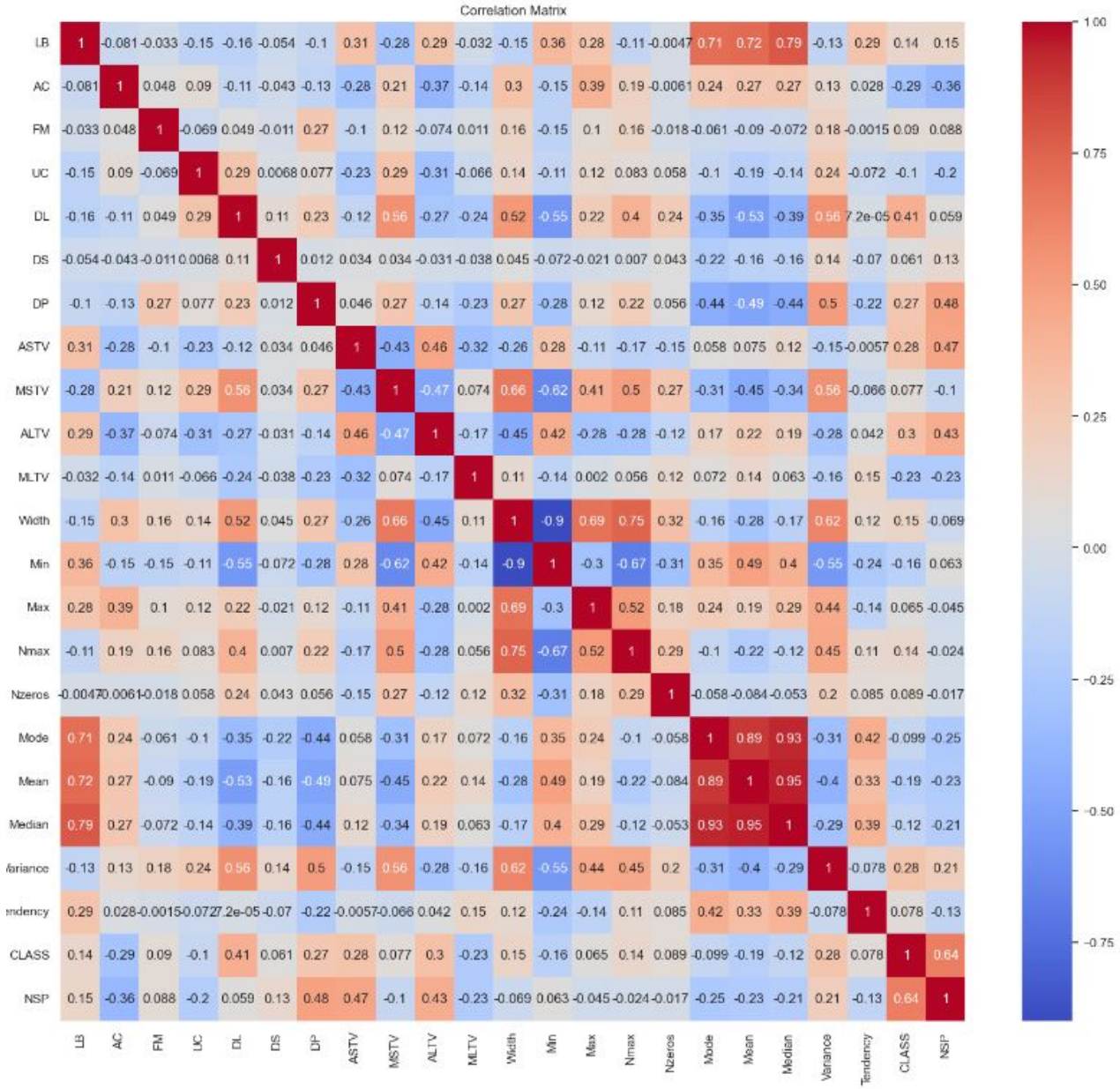


Figure 4.4: Correlation Matrix, Cardiotocography Dataset

Heart Disease Preprocessing

In the Heart Disease dataset, a well-optimized dataset was encountered, where minimal preprocessing was required. Which only had to handle a few missing values by removing them from the dataset.

However, one interesting aspect of the dataset was the variable to be predicted. Originally, it contained values of 0, 1, 2, 3, and 4. Upon closer examination, we discovered that the values 1, 2, 3, and 4 were actually intended to represent the same underlying concept, as specified by the dataset creator. To simplify the prediction task and reduce complexity, we decided to merge the values 2, 3, and 4 into a single category and replaced them with 1. This allowed us to convert the variable into a binary classification problem, with two distinct values to predict, heart disease or not.

By making this adjustment, a more straightforward and streamlined prediction task was achieved, eliminating the need to differentiate between multiple categories that represented the same outcome. This simplification not only enhanced the interpretability of the results but also reduced potential confusion and improved the overall performance of the machine learning models applied to the dataset.

As seen below in Figure 4.5, the variable to be predicted in the dataset, namely the presence or absence of heart disease, exhibited a well-balanced distribution. This means that the instances with and without heart disease were relatively evenly distributed within the dataset. Such balanced distribution is beneficial in machine learning tasks as it helps prevent bias towards any specific outcome during model training and evaluation.

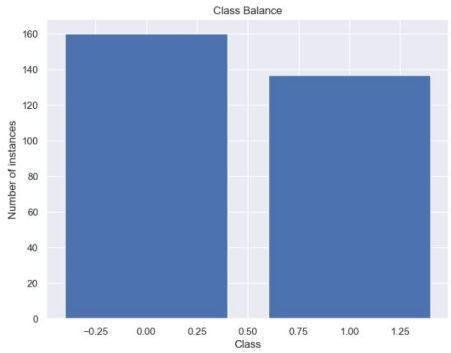


Figure 4.5: Variable to predict, Heart Disease

The Heart Disease dataset's correlation matrix (Figure 4.6) demonstrates a significant inverse association between a number of factors. A useful tool for comprehending the relationships and interactions between various variables in a dataset is the correlation matrix. In this instance, it implies that shifts in one variable are frequently accompanied by shifts in a different variable in the opposite direction.

In figure Figure 4.6, it can be seen that although most of the variables do not have a clear correlation, in fact, most of them present a slight inverse correlation, the variables "age" and "cp" stand out with "thalach" which present an inverse correlation of -0.39 and -0.34 respectively and additionally "exang", "oldpeak", "slope", "ca", "thal" and the variable to be predicted "num" all present an inverse relationship above -0.25

The dataset's significant inverse correlations show that some variables have a propensity to vary in the opposite ways. Due to the potential for redundant information provided by highly correlated variables, this information can be helpful in feature selection or dimensionality reduction.

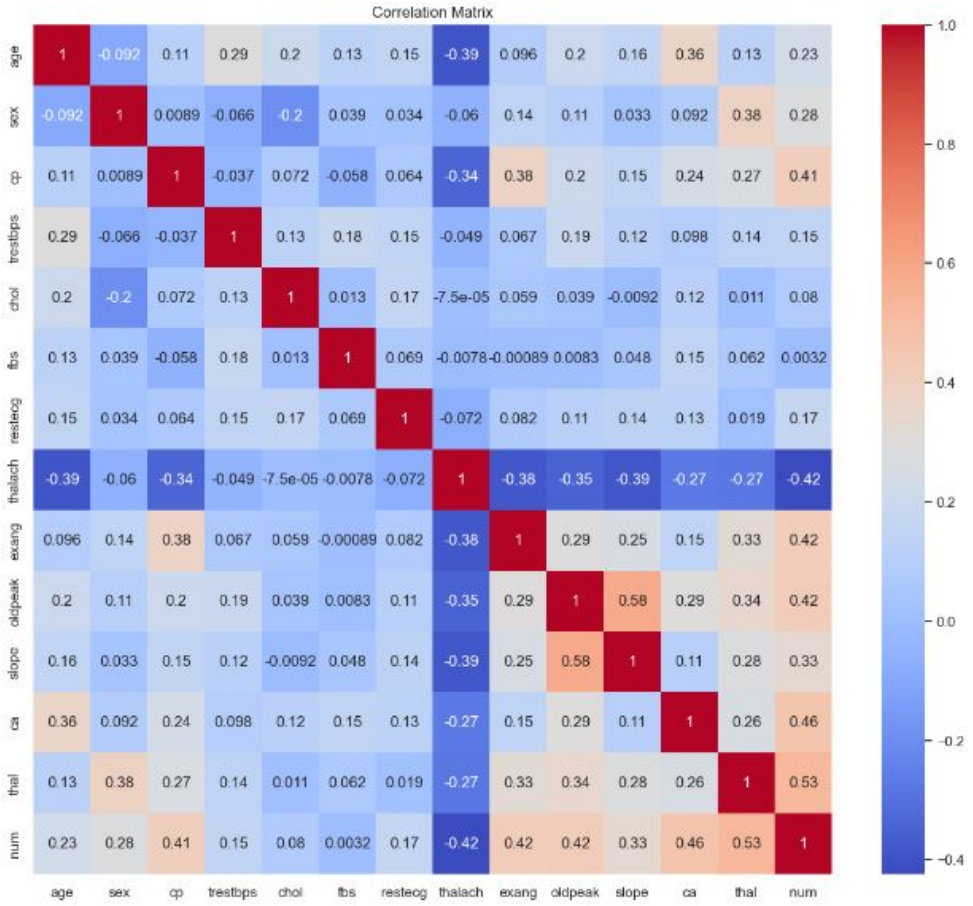


Figure 4.6: Correlation Matrix, Heart Disease Dataset

4.2 Results

The results section presents the findings and outcomes of the studies conducted using the datasets. Through extensive analysis and application of machine learning techniques, we gained valuable insights into the predictive capabilities of the models developed. These insights are crucial for understanding the performance and effectiveness of the machine learning system in predicting patient illnesses based on demographic and laboratory data.

The studies involved evaluating the accuracy, precision, recall, and other relevant performance metrics of the models. The results demonstrate the ability of the developed system to make accurate predictions, providing valuable information for healthcare professionals and decision-makers. Additionally, the studies examined the generalizability of the models by assessing their performance on unseen data and validating their robustness.

EMRBots Results

In the EMRBots dataset, there were thousands of variables to predict initially, which posed a significant challenge for the model's effectiveness. As described in the preprocessing section, we made the decision to group and generalize these diseases in order to increase the precision of our predictions.

However, the results obtained were not optimal, as attempting to predict multiple complex diseases solely based on demographic and laboratory data proved to be quite challenging. Nonetheless, this setback served as motivation to focus specifically on heart diseases and develop the capability to predict them accurately.

Decision Tree Classifier Results

The Decision Tree Classifier algorithm yielded relatively low accuracy and F1-score macro values, with an **accuracy of 24% and an F1-score macro of 19%**. These results indicate that the model's performance in classifying the target variable was not satisfactory. Several factors might have contributed to this outcome.

Firstly, the complexity and nature of the dataset might have posed challenges for the Decision Tree Classifier algorithm. If the dataset contained a large number of features or had high-dimensional data, it could have led to overfitting, where the model becomes too specific to the training data and fails to generalize well to new, unseen data. This can result in poor accuracy and limited predictive power.

Additionally, the decision tree algorithm is known to have limitations when dealing with imbalanced datasets, where the distribution of target classes is uneven. If the dataset

had imbalanced classes, with one or more classes being significantly more prevalent than others, it could have affected the model's ability to learn and accurately predict the minority class(es).

Random Forest Classifier Results

Despite initial expectations that the Random Forest Classifier algorithm would yield better results, it produced relatively low accuracy and F1-score macro values, with an **accuracy of 33% and an F1-score macro of 15%**. Several factors may have contributed to these outcomes.

Firstly, similar to the Decision Tree Classifier, the complexity of the dataset and the presence of a large number of features could have affected the Random Forest Classifier's performance. If the dataset contained irrelevant or noisy features, it could have led to overfitting or reduced the algorithm's ability to capture the true underlying patterns and make accurate predictions.

Furthermore, the Random Forest algorithm relies on the principle of ensemble learning, where multiple decision trees are trained on different subsets of the data and their predictions are aggregated.

Support Vector Machine (SVM) Results

Despite high hopes for the Support Vector Machine (SVM) algorithm as our last resort, it also yielded relatively low accuracy and F1-score macro values, with an **accuracy of 34% and an F1-score macro of 10%**. Several factors could have contributed to these results.

Also, despite the utilization of GridSearch [37] to select the best hyperparameters for the Support Vector Machine (SVM) algorithm and the algorithms mentioned above, the obtained results still exhibited relatively low accuracy and F1-score macro values. One possible reason is the complexity and non-linearity of the dataset. SVM performs well when the data is linearly separable or can be transformed into a higher-dimensional space where it becomes linearly separable. However, if the dataset contains complex relationships or overlaps between classes, SVM may struggle to find an optimal decision boundary and accurately classify the data.

Furthermore, the performance of SVM can be highly dependent on the chosen kernel function. While GridSearch helps in finding the optimal combination of hyperparameters, it cannot guarantee the selection of the most appropriate kernel for a particular dataset.

Cardiotocography Results

Unlike the EMRBots dataset, the Cardiotocography dataset showed significant improvement in prediction accuracy. This improvement can be attributed to a different approach taken in this study, focusing on a single disease: the presence or suspicion of fetal cardiac disease.

By targeting a single disease, we were able to train our machine learning models specifically for the classification task of identifying fetal cardiac disease. This narrower focus allowed the models to learn patterns and features relevant to this particular condition, resulting in improved prediction performance. Additionally, the dataset exhibited a well-balanced distribution of the target variable, which further facilitated accurate classification.

In this study, the attempt to generate a Receiver Operating Characteristic (ROC) curve [38] was unsuccessful due to encountering the error message"*ValueError: multiclass format is not supported*" This error arises because the ROC curve is primarily designed for binary classification tasks or evaluating the performance of binary classifiers.

Decision Tree Classifier Results

The Decision Tree Classifier algorithm in the Cardiotocography dataset achieved exceptional **accuracy and F1-score macro values of 98% and 97%**, respectively, which are remarkable results for a predictive model.

However, a deeper analysis revealed a potential issue. The learning curve [39] plot demonstrated a high bias, indicating that the model was overfitting the training data. This overfitting was evident from the consistent 100% training score across different training set sizes as seen in the Figure 4.7a.

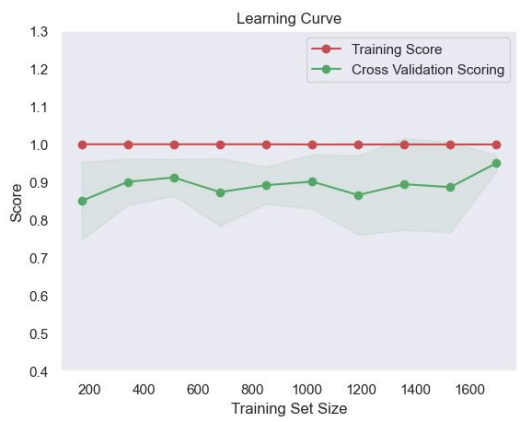
While achieving such high accuracy on the training data may seem promising, it indicates that the model is simply memorizing the training examples and may not generalize well to unseen data. In real-world scenarios, where the model encounters new cases, this overfitting can lead to poor performance. The model may struggle to adapt and make accurate predictions for unseen fetal health instances.

Random Forest Classifier Results

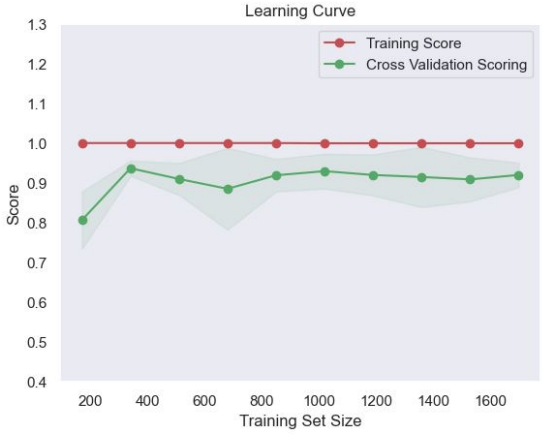
The Decision Tree Classifier algorithm in the Cardiotocography study achieved an impressive **accuracy of 98% and a macro F1-score of 97%**, which are exceptional values indicating a high level of predictive performance. However, same as in the previous model, a crucial issue was observed when analyzing the learning curve of the model.

The learning curve, as seen in the Figure 4.7b, depicts the performance of the model as

a function of the training set size. In this case, it was observed that the model exhibited a high bias, as indicated by the consistently high training score of 100% across different training set sizes. This suggests that the model was overfitting the training data, essentially memorizing it instead of generalizing well to unseen data.



(a) Learning Curve of Decision Tree Classifier from Cardiotocography Dataset



(b) Learning Curve of Random Forest Classifier from Cardiotocography Dataset

Figure 4.7: Learning Curve of Decision Tree & Random Forest Classifier from Cardiotocography Dataset

Support Vector Machine Results

The Support Vector Machine (SVM) algorithm in the Cardiotocography study achieved an **accuracy of 97% and a macro F1-score of 92%**, which are slightly lower than the previous models. However, a significant advantage of the SVM model is its excellent learning curve, indicating that the model is not memorizing the training data.

The learning curve of the SVM model shows a consistent increase in both the training and cross-validation scores as the training set size increases, as seen in the Figure 4.8. This suggests that the model is effectively learning from the data and improving its performance with more training examples.

This robust learning curve indicates that the SVM model has the potential to perform well on new, unseen instances, making it a reliable tool for predicting fetal health status based on the Cardiotocography dataset. Although its accuracy and F1-score may be slightly lower compared to other models, the SVM's ability to generalize and avoid overfitting makes it a valuable choice for practical applications.

Finally, since this is the learning model that best suits our standards, the confusion matrix is presented, where the following relationship can be seen, as seen in Figure 4.9:

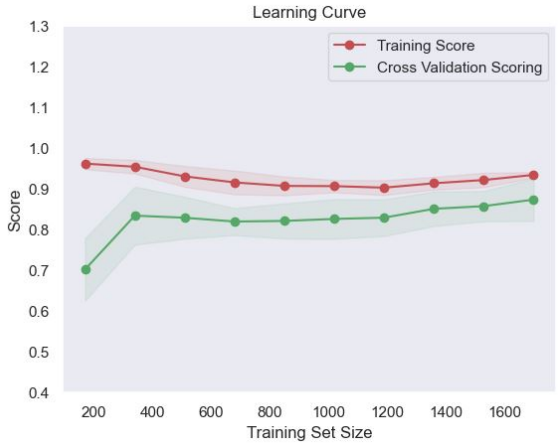


Figure 4.8: Learning Curve of Support Vector Machine from Cardiotocography Dataset

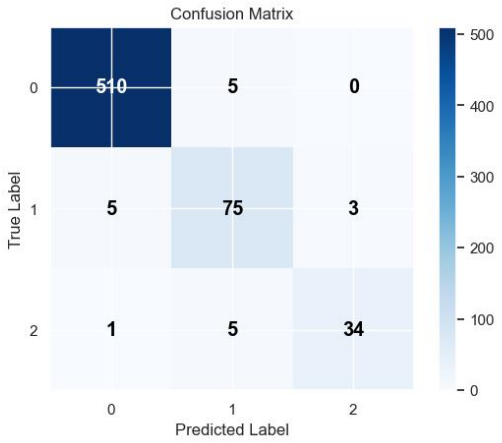


Figure 4.9: Confusion Matrix from Cardiotocography Dataset

- For Class 1, there are 510 true positives and 5 false negatives.
- For Class 2, there are 75 true positives, 5 false positives, and 3 false negatives.
- For Class 3, there are 34 true positives, 1 false positive, and 5 false negatives.

In summary, the confusion matrix shows the number of samples correctly classified (true positives) and incorrectly classified (false positives and false negatives) for each class. This provides insights into the performance of the classification model and helps evaluate its accuracy in predicting each class.

Heart Disease Results

In the Heart Disease dataset, similar to the Cardiotocography dataset, we focused on predicting the presence or absence of heart disease in adults based on a set of relevant variables.

Decision Tree Classifier Results

The decision tree classifier applied to the Heart Disease dataset achieved an **accuracy of 83% and a macro F1-score of 83%**, indicating decent performance in predicting the presence of heart disease. However, same as last dataset, a closer analysis of the learning curve reveals some limitations, as seen in the Figure 4.10a. The curve demonstrates a high bias in the cross-validation scoring, and the training score remains consistently at 100%.

The high training score suggests that the decision tree classifier is overfitting the training data, effectively memorizing the patterns and details specific to the training set. This memorization-based approach may not generalize well to unseen data and real-world scenarios, leading to poor performance in practical applications.

Random Forest Classifier Results

The random forest classifier applied to the Heart Disease dataset exhibited similar performance to the decision tree classifier mentioned earlier. It achieved an **accuracy of 83% and a macro F1-score of 83%**, which are comparable to the results obtained from the decision tree classifier. However, the learning curve analysis revealed the same limitation observed in the decision tree classifier, as shown in the Figure 4.10b.

(a) Learning Curve of Decision Tree Classifier from Heart Disease Dataset

(b) Learning Curve of Random Forest Classifier from Heart Disease Dataset

Figure 4.10: Learning Curve of Decision Tree & Random Forest Classifier from Heart Disease Dataset

Support Vector Machine Results

The support vector machine (SVM) algorithm applied to the Heart Disease dataset outperformed both the decision tree classifier and the random forest classifier. It achieved an **accuracy of 85% and a macro F1-score of 83%**, surpassing the performance of the other two algorithms. Notably, the SVM model exhibited an excellent learning curve, indicating that it is not memorizing the training data.

Figure 4.11: Learning Curve of Support Vector Machine from Heart Disease Dataset

The learning curve analysis, shown in Figure 4.11, demonstrated that the SVM model has a reasonable balance between bias and variance. The training and cross-validation scores converge to a stable and high level of performance, suggesting that the model is effectively capturing the underlying patterns and generalizing well to unseen data. This characteristic makes the SVM classifier a reliable choice for predicting heart disease based on the given dataset.

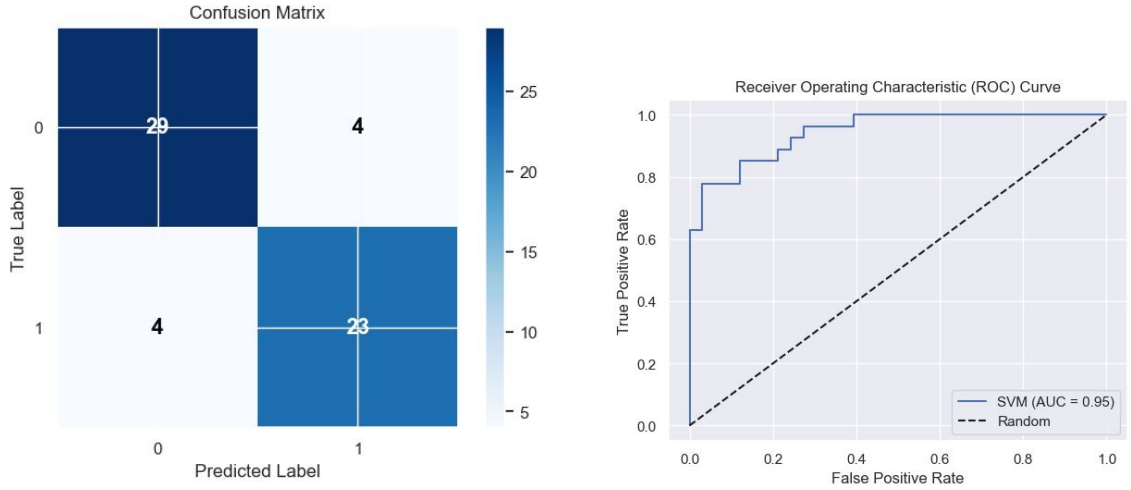
Finally, since this is the learning model that best suits our standards, the confusion matrix is presented, where the following relationship can be seen, as seen in Figure 4.12a:

- For class 1, there are 29 correct predictions (true positives) and 4 incorrect predictions as class 2 (false negatives).
- For class 2, there are 23 correct predictions (true positives) and 4 incorrect predictions as class 1 (false positives).

In addition to the high accuracy and F1-score, the support vector machine (SVM) classifier applied to the Heart Disease dataset also exhibits a remarkable receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate against the false

positive rate at various classification thresholds, providing valuable insights into the model's discriminatory power.

In this study, as shown in the Figure 4.12b, the SVM model yields a well-shaped ROC curve with an area under the curve (AUC) of 0.95. The AUC is a widely used metric to assess the overall performance of a binary classifier, where a value closer to 1 indicates better discrimination between the positive and negative classes. In this case, the AUC value of 0.95 indicates that the SVM model has a high probability of assigning higher scores to instances with heart disease compared to those without. The high AUC value reinforces the reliability and discriminative ability of the SVM model, further supporting its suitability for predictive healthcare applications.



(a) Confusion Matrix from Heart Disease Dataset

(b) ROC Curve from Heart Disease Dataset

Figure 4.12: Confusion Matrix & ROC Curve from Heart Disease Dataset

F1-score Summary

The F1-score, as shown in table 4.1, is a crucial metric when evaluating a prediction model because it considers both precision and recall simultaneously, providing a balanced measure of the model's performance in terms of both correctly identifying positive instances and minimizing false positives and false negatives.

	Decision Tree	Random Forest	Support Vector Machine
EMRBots	0.19	0.15	0.10
Cardiotocography	0.97	0.97	0.92
Heart Disease	0.83	0.83	0.83

Table 4.1: F1-score Summary

4.3 Analysis

In this section, the main focus is on examining the key variables employed in the predictive modeling strategy and comparing them with the essential variables used in medical diagnosis for disease detection. The aim is to explore the correlation between predictive modeling methods and the clinical significance of specific variables, considering both perspectives.

A comprehensive analysis will be conducted to compare the variables identified as influential in the predictive models with the variables commonly used in medical diagnosis. This analysis will provide insights into the alignment between predictive modeling techniques and the clinical importance of variables in disease detection. The examination will shed light on the strengths and limitations of the approach, contributing to a deeper understanding of the relationship between data-driven predictions and traditional medical diagnostic practices.

Throughout this section, the focus will be on contrasting the identified variables and studying their significance, aiming to elucidate the synergies and disparities between predictive modeling and medical diagnosis. The objective is to provide a comprehensive perspective on the relationship between these two approaches and enhance our understanding of the role of variables in disease detection.

Based on the analysis and comparison of the different classification algorithms used in the study, it has been determined that the Support Vector Machine (SVM) algorithm has shown the most promising results in terms of accuracy and F1-score. Therefore, for further analysis and interpretation, the focus will be solely on the SVM algorithm.

EMRBots Analysis

In the case of the EMRBots dataset, the analysis revealed that the most influential variables for predicting diseases were related to the marital status of the patients, specifically White, Divorced, Married, Separated, and Single, as shown in Figure 4.13. Interestingly, these variables do not possess direct medical significance, at least based on current medical

knowledge. This finding suggests that the dataset may contain confounding factors or biases that have led to these variables appearing significant in the predictive modeling process.

On the other hand, the laboratory variables in the EMRBots dataset did not demonstrate a substantial impact on disease prediction. This implies that the traditional laboratory measurements included in the dataset may not carry significant weight in the predictive model. It is important to note that these findings do not imply that laboratory variables are irrelevant in the medical context but rather suggest that, within the specific context of this dataset, they did not contribute significantly to the predictive power of the model.

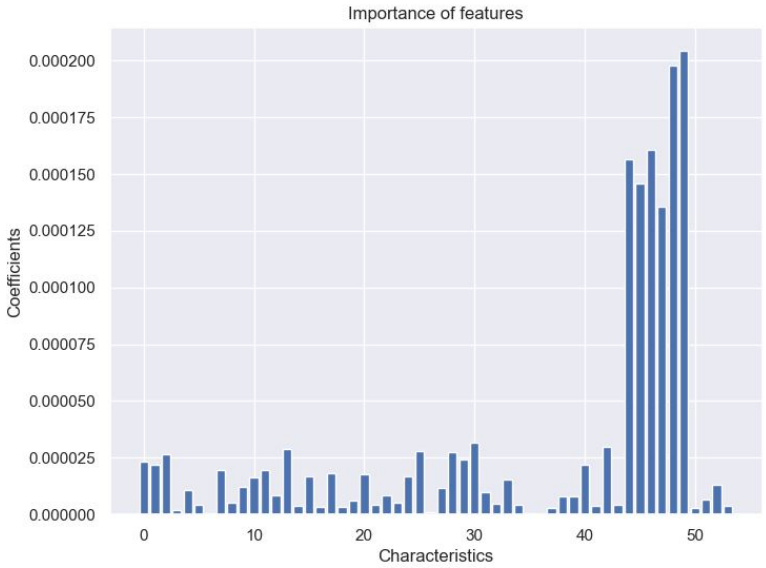


Figure 4.13: Importance of Features from EMRBots Dataset

As a result, the inability to achieve a highly accurate predictive model may be attributed to the inclusion of non-medically relevant variables and the limited impact of laboratory measurements. The presence of variables related to marital status, without a clear medical explanation, may introduce noise and hinder the model's ability to accurately predict diseases. These observations highlight the importance of critically evaluating the relevance and appropriateness of variables used in predictive modeling and considering domain-specific knowledge when interpreting the results.

Cardiotocography Analysis

In the Cardiotocography dataset, the analysis revealed that the most significant variables for predicting the presence or absence of fetal heart disease were AC (accelerations of the fetal heart rate per second), Mean, and Mode of the fetal heart rate histogram, as shown in Figure 4.14. These variables are directly related to the fetal heart activity and can provide

valuable insights into the health status of the fetus.

From a medical perspective, the relevance of these variables aligns with the detection of heart diseases in fetuses [40]. The AC variable reflects the accelerations of the fetal heart rate, which can indicate a healthy cardiac response. Abnormalities or deviations in AC values may indicate potential heart issues. Similarly, the Mean and Mode of the fetal heart rate histogram provide valuable information about the distribution and patterns of the fetal heart rate, which can help identify irregularities or abnormalities associated with heart diseases.

The correlation between these variables and the detection of fetal heart diseases in real-life medical practice demonstrates the relevance and importance of these variables in assessing the cardiac health of the fetus. By leveraging these variables in predictive modeling, we can potentially develop a reliable system for predicting fetal heart diseases based on Cardiotocography data.

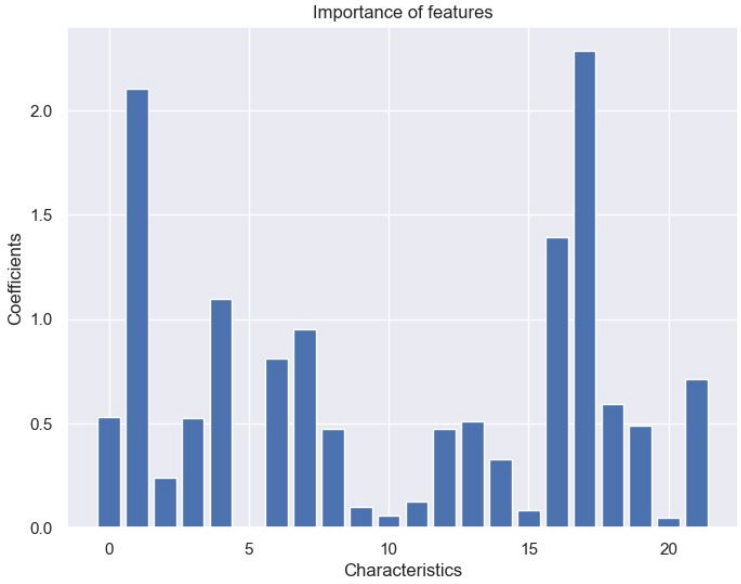


Figure 4.14: Importance of Features from Cardiotocography Dataset

Heart Disease Analysis

The variable "ca" in the Heart Disease dataset, which represents the number of major blood vessels colored by fluoroscopy (ranging from 0 to 3), and the variable"cp", which represents the type of chest pain (typical angina, atypical angina, non-anginal pain, or asymptomatic), were found to be the most influential in predicting the presence of heart disease, as shown in Figure 4.15.

In the context of real-world medical detection of heart diseases, the significance of these variables aligns with clinical practice [41]. The number of major blood vessels colored by fluoroscopy is a crucial indicator as it reflects the extent of blockages or narrowing in the coronary arteries. The presence of multiple diseased vessels indicates a higher likelihood of significant coronary artery disease. Additionally, the type of chest pain is a valuable diagnostic clue in assessing the nature and severity of cardiac conditions. Typical angina, characterized by chest pain triggered by physical exertion and relieved by rest or nitroglycerin, often points towards coronary artery disease.

Therefore, the importance of these variables in predicting heart disease aligns with their clinical relevance in the medical field. By leveraging these key indicators, predictive models can effectively identify individuals at higher risk of heart disease, assisting healthcare professionals in making informed decisions and providing appropriate interventions for patients.

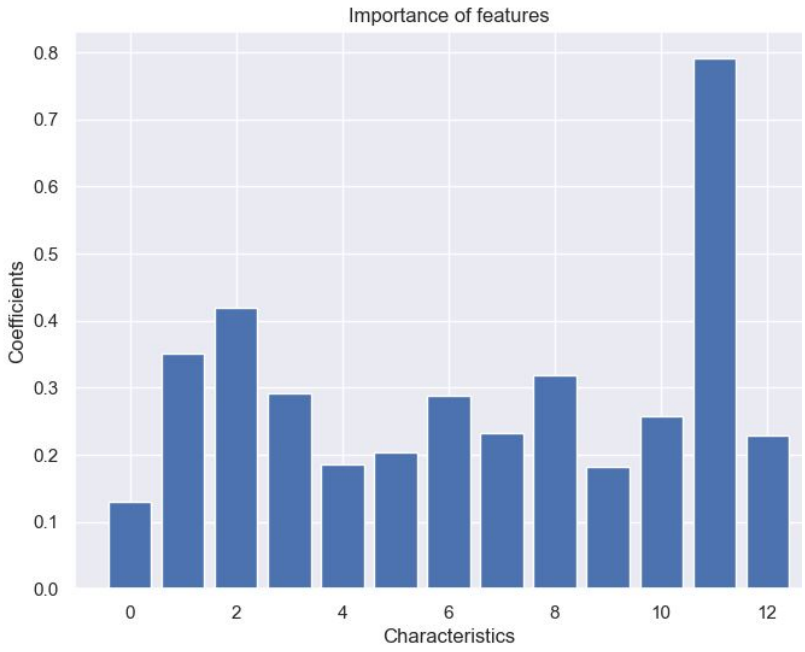


Figure 4.15: Importance of Features from Heart Disease Dataset

CHAPTER 5

Conclusions and future work

5.1 Conclusions

In conclusion, the application of machine learning techniques in healthcare holds tremendous potential for transforming the field of medicine. Through the analysis of diverse datasets and the utilization of preprocessing techniques, valuable insights can be derived to aid in disease prediction, patient diagnosis, and treatment personalization. The findings from this study shed light on the strengths and limitations of machine learning models in healthcare and highlight the importance of variables in disease detection.

The investigation encompassed three distinct datasets: EMRBots, Cardiotocography, and Heart Disease. Each dataset presented unique challenges and opportunities. In the case of EMRBots, the presence of numerous variables initially hindered the effectiveness of the models in predicting diseases. However, through the grouping and generalization of diseases, some improvement in prediction precision was achieved, albeit falling short of expectations. This emphasized the complexity of predicting multiple diseases solely based on demographic and laboratory data.

Conversely, the Cardiotocography dataset exhibited more promising results. By focusing specifically on the prediction of fetal heart diseases, significant variables such as heart rate

accelerations (AC) and features of the fetal heart rate histogram emerged as key predictors. These variables aligned closely with established medical indicators used in real-world detection of fetal heart conditions. This underscores the potential for machine learning models to complement existing medical practices and provide valuable insights for accurate disease detection.

Similarly, the Heart Disease dataset enabled a focused analysis on predicting cardiac conditions in adults. Variables such as the number of major blood vessels colored by fluoroscopy (CA) and the type of chest pain (CP) emerged as influential factors in disease prediction. Importantly, these variables closely correlated with recognized medical indicators used in the diagnosis of heart diseases. The incorporation of such variables into the models yielded reasonable accuracy in predicting the presence or absence of heart disease.

It is important to acknowledge that the effectiveness of machine learning models in disease prediction is influenced by several factors, including data quality, feature selection, and algorithm choice. Preprocessing techniques, such as data cleaning, normalization, and feature engineering, play a critical role in enhancing model performance. Moreover, careful consideration should be given to the selection of the classification algorithm, as different algorithms may yield varying results based on dataset characteristics and the nature of the problem at hand.

In summary, the application of machine learning in healthcare has the potential to revolutionize medical diagnosis and treatment. This study provides valuable insights into the interplay between predictive modeling approaches and established medical practices. By combining the power of machine learning algorithms with clinical expertise, it is possible to pave the way for more accurate disease detection, personalized medicine, and improved patient outcomes. As technology advances and more comprehensive datasets become available, the future of machine learning in healthcare appears promising, promising a transformative impact on medical challenges and ultimately enhancing the delivery of healthcare services.

5.2 Achieved goals

The objectives proposed in section 1.2 were successfully completed during the study, going beyond what was proposed:

1. Obtained complete datasets containing real values on physiological, demographic, or analytical variables and performed in-depth analysis and preprocessing of the data. The datasets of Cardiotocography, EMRBots, and Heart Disease were thoroughly

examined, ensuring the availability of comprehensive information on various variables. The researchers employed advanced preprocessing techniques, including handling missing values, categorical variable encoding, and data standardization using StandardScaler. These steps ensured the suitability and quality of the datasets for further analysis and modeling.

2. Developed and evaluated multiple prediction models using different datasets, aiming to achieve an f1-macro accuracy above 80 percent. Several machine learning algorithms, including Decision Tree Classifier, Random Forest Classifier, and Support Vector Machine (SVM), were implemented. Performance evaluation metrics such as accuracy, precision, recall, and f1-score were utilized to assess the models' effectiveness. Cross-validation techniques and hyperparameter tuning using GridSearchCV were applied to optimize the models' performance.

3. Gained insights into the physiological significance of the models' key characteristics, enhancing understanding of underlying patterns and relationships. Through feature importance analysis, we identified the variables that significantly influenced the models' predictions. In the Cardiotocography dataset, the variables related to heart accelerations (AC) and features of the fetal heart histogram (mean and mode) were found to be crucial in predicting the presence of cardiac disease in the fetus. Similarly, in the Heart Disease dataset, variables such as the number of colored vessels (CA) and the type of chest pain (CP) were identified as influential factors in detecting heart disease. The researchers drew connections between these key variables and their clinical importance in real-world disease detection scenarios, demonstrating the relevance and reliability of the predictive models.

5.3 Future work

Future research in the areas of healthcare and machine learning can take a number of different directions. These guidelines are intended to improve the precision, applicability, and usefulness of prediction models in actual healthcare settings.

First off, adding new data sources and broadening the range of factors can help forecast outcomes and give a more thorough picture of illness trends. Integrating genetic, environmental, lifestyle, and patient-generated data may offer important new perspectives on the development and course of disease. The predictive models may be able to capture more complex interactions and perform better overall by using a larger variety of factors.

Second, the quality and relevance of the input data can be improved by improving preprocessing procedures and feature engineering approaches. The most informative features can be found, and the models' effectiveness can be increased, by investigating advanced feature selection algorithms and dimensionality reduction approaches. Additionally, addressing issues like missing data, unbalanced classes, and outlier detection through the development of unique data preparation methods tailored to healthcare datasets might result in more robust and trustworthy predictive models.

Additionally, the incorporation of explainable AI (XAI) approaches helps make machine learning model predictions transparent and understandable. Healthcare practitioners can develop trust and confidence in the model outputs by comprehending the underlying variables and patterns that contribute to the forecasts. This can then make it easier for machine learning to be included into healthcare decision-making procedures.

The assessment and validation of the produced models in actual clinical settings is a crucial subject that requires more investigation. Large-scale clinical studies or the implementation of the models in healthcare systems might provide important details about their efficiency, effect on patient outcomes, and potential drawbacks. It is possible to guarantee the generalizability and dependability of the models in real-world settings by evaluating their performance on a variety of patient groups and healthcare contexts.

In addition, current developments in machine learning and artificial intelligence approaches call for continued study and investigation. Emerging techniques like deep learning, reinforcement learning, and ensemble methods present chances for enhanced prognostic accuracy and the identification of intricate linkages in medical data. The performance and application of these cutting-edge techniques in healthcare datasets could be examined in order to improve the models' predictive powers.

Finally, resolving ethical, privacy, and security issues related to the use of sensitive patient data is a crucial component of future work. For machine learning models to be successfully implemented and accepted in the healthcare industry, it will be essential to create solid frameworks for data governance, protect patient privacy, and ensure data security.

APPENDIX A

Impact of this project

In this appendix we will analyze the different impacts in relation to this project, including social, economic, environmental and ethical impacts.

A.1 Social impact

A considerable social impact may result from the use of predictive modeling for illness identification based on demographic and laboratory data. Healthcare professionals can intervene at a critical juncture by precisely identifying people who are at high risk of contracting specific diseases, which may improve treatment outcomes and save costs. By encouraging preventative actions and targeted interventions, this can improve the general health and wellbeing of the populace. Additionally, by enabling more equitable access to individualized care, regardless of socioeconomic variables, the adoption of data-driven initiatives can aid in addressing healthcare inequities.

A.2 Economic impact

Predictive modeling's use in disease identification has potential to have a significant economic impact. Healthcare systems can more effectively deploy resources, decreasing unnecessary medical treatments, and maximizing healthcare costs by precisely identifying persons at risk. By reducing the need for costly therapies linked to severe illness stages, early detection and intervention can also lead to cost savings. Additionally, the application of data-driven methods in healthcare can promote technological development, generating new employment possibilities and promoting economic expansion in the healthcare and technology industries.

A.3 Environmental impact

Predictive modeling for illness diagnosis can indirectly support sustainability initiatives even though its effects on the environment might not be immediately apparent. Predictive modeling can assist lower the total demand for medical resources, including energy-intensive procedures and interventions, by increasing diagnostic accuracy and optimizing the allocation of healthcare resources. Additionally, it is feasible to lessen the strain on healthcare systems by encouraging preventative treatment and early identification, which can result in a more effective use of resources and a less environmental impact connected with healthcare-related activities.

A.4 Ethical impact

Predictive modeling for disease identification has important ethical ramifications that need to be carefully considered. It is crucial to make sure that data is collected, stored, and used in a way that respects personal privacy and confidentiality. To build trust and prevent biases or discrimination, transparency in the modeling process, including the choice of variables and algorithms, is essential. Additionally, the ethical and responsible application of predictive models should take into account any potential repercussions and guarantee that decisions are made in the best interests of patients and society at large. In order to handle potential moral conundrums like the equitable distribution of healthcare resources and the appropriate use of individual health information, safeguards must be in place.

APPENDIX B

Economic budget

In ths appendix we will analyse the economic budget in relation to this project, including project structure, physical resources, human resources and taxes.

B.1 Physical resources

The physical resources utilized for the project include a computer system with the following specifications:

- **CPU:** AMD Ryzen 5 5600X 6-Core Processor
- **GPU:** Nvidia RTX 3060 Ti 8GB GDDR6X
- **RAM:** 16GB DDR4 3200 MHz
- **Storage:** 1TB HDD + 1TB M.2 SSD + 240GB SSD

The estimated cost of the hardware is around 1200$.

B.2 Project structure

The project was planned with a number of different tasks in mind, having the days structured as it is described in the text below:

Activity	**Days**
Researching and Learning new concepts	48
Learning Technologies Required (Machine Learning, Pandas, etc.)	72
Planning the structure	15
Development and Understanding of the Algorithm	90
Writing the final document	15
Total	240

B.3 Human Resources

The necessary budget to cover the cost of human resources would be for one person, taking into account that the project was made individually and estimating a salary of 2,000$ per month, that would be 66.66$ per day. Taking into consideration that the project took 240 days it would have a cost of 16,000$ (without taxes).

B.4 Taxes

If the product was selled locally, it would be subjected to the Value-Added Tax in Spain that is 21% of the product value. Meaning the final cost of the project is valued at **19,359$**

Bibliography

[1] Qlik. How do big data and ai work together? https://www.qlik.com/us/augmented-analytics/big-data-ai, 2020. Accessed: 2023-05-08.

[2] Sharvari Shukla Nishita Mehta, Anil Pandit. Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. https://www.sciencedirect.com/science/article/pii/S1532046419302308, 2019. Accessed: 2023-05-08.

[3] Maya Hamdanieh Louna Ftouni Racha Ftouni, Baraa AlJardali and Nariman Salem. Challenges of telemedicine during the covid-19 pandemic: a systematic review. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9351100/, 2022. Accessed: 2023-05-08.

[4] Python Software Foundation. What is python? executive summary. https://www.python.org/doc/essays/blurb/, 2016. Accessed: 2023-05-10.

[5] Datacamp. Python in healthcare: Ai applications in hospitals. https://www.datacamp.com/blog/python-in-healthcare-ai-applications-in-hospitals, 2022. Accessed: 2023-05-10.

[6] Senerath Mudalige Don Alexis Chinthaka Jayatilake and Gamage Upeksha Ganegoda. Involvement of machine learning tools in healthcare decision making. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7857908/, 2021. Accessed: 2023-05-10.

[7] Jupyter Team. Jupyter project documentation. https://docs.jupyter.org/en/latest/, 2015. Accessed: 2023-05-09.

[8] Scikit-Learn developers. Scikit-learn tutorials. https://scikit-learn.org/stable/tutorial/index.html, 2023. Accessed: 2023-05-09.

[9] Gunjan Goyal. Application of machine learning in medical domain! https://www.analyticsvidhya.com/blog/2021/06/application-of-machine-learning-in-medical-domain/, 2021. Accessed: 2023-05-09.

[10] ActiveState. What is pandas in python? everything you need to know. https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/, 2022. Accessed: 2023-05-10.

[11] NumPy. What is numpy? https://numpy.org/doc/stable/user/whatisnumpy.html, note = Accessed: 2023-05-10, 2014.

[12] ActiveState. What is pyplot in matplotlib? https://www.activestate.com/resources/quick-reads/what-is-pyplot-in-matplotlib/, 2022. Accessed: 2023-05-11.

[13] Margaret Rouse. What does data set mean? https://www.techopedia.com/definition/3348/data-set-ibm-mainframe, 2022. Accessed: 2023-05-09.

[14] Uri Kartoun. Emrbots. http://www.emrbots.org/, 2018. Accessed: 2023-05-09.

[15] Cao Xiao Tengfei Ma and Fei Wang. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. https://epubs.siam.org/doi/10.1137/1.9781611975321.30, 2018. Accessed: 2023-05-21.

[16] World Health Organization. International classification of diseases (icd). https://icd.who.int/browse10/2019/en, 2019. Accessed: 2023-05-09.

[17] Machine Learning Repository. Cardiotocography data set. https://archive.ics.uci.edu/ml/datasets/cardiotocography, 2010. Accessed: 2023-05-09.

[18] S. Nickolas TM. Ramla, S. Sangeetha. Fetal health state monitoring using decision tree classifier from cardiotocography measurements. https://ieeexplore.ieee.org/abstract/document/8663047?casa_token=Uxwn11Fux5gAAAAA:XDNwM-565We92dmZZzcZFQZ1Awlo6kLkgukFnjPSAbskz_qBq5RCmsCAbY-bbxx_Sind9bGg5g, 2018. Accessed: 2023-05-21.

[19] Machine Learning Repository. Heart disease data set. https://archive.ics.uci.edu/ml/datasets/heart+disease, 1988. Accessed: 2023-05-09.

[20] Midhun Chakkravarthy Tai-hoon Kim Bhanu Prakash Doppala, Debnath Bhattacharyya. A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. https://link.springer.com/article/10.1007/s10619-021-07329-y, 2023. Accessed: 2023-05-21.

[21] Craig Stedman. Why data quality is important? https://www.techtarget.com/searchdatamanagement/definition/data-quality, 2021. Accessed: 2023-05-23.

[22] Ethem Alpaydin. *Machine learning*. Mit Press, 2021.

[23] Nagesh Singh Chauhan. Decision tree algorithm, explained. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html, 2022. Accessed: 2023-05-25.

[24] Scikit-Learn. sklearn.tree.decisiontreeclassifier. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html, 2023. Accessed: 2023-05-25.

[25] Scikit-Learn. sklearn.ensemble.randomforestclassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, 2023. Accessed: 2023-05-27.

[26] Tony Yiu. Understanding random forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2, 2019. Accessed: 2023-05-27.

[27] Scikit-Learn. Support vector machines. https://scikit-learn.org/stable/modules/svm.html, 2023. Accessed: 2023-05-27.

[28] IBM. How svm works. https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works, 2021. Accessed: 2023-05-28.

[29] ICD-10. International statistical classification of diseases and related health problems 10th revision. https://icd.who.int/browse10/2019/en, 2019. Accessed: 2023-05-24.

[30] Pandas. pandas.dataframe.merge. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html, 2023. Accessed: 2023-05-24.

[31] Pandas. pandas.dataframe.join. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.join.html, 2023. Accessed: 2023-05-24.

[32] Pandas. pandas.dataframe.pivot. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.pivot.html, 2023. Accessed: 2023-05-24.

[33] Scikit-Learn. sklearn.preprocessing.onehotencoder. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html, 2023. Accessed: 2023-05-24.

[34] Pandas. pandas.concat. https://pandas.pydata.org/docs/reference/api/pandas.concat.html, 2023. Accessed: 2023-05-24.

[35] Scikit-Learn. sklearn.preprocessing.standardscaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, 2023. Accessed: 2023-05-24.

[36] CFI Team. Correlation matrix. https://corporatefinanceinstitute.com/resources/excel/correlation-matrix/#:~:text=A%20correlation%20matrix%20is%20simply,patterns%20in%20the%20given%20data., 2023. Accessed: 2023-05-24.

[37] Farhad Malik. What is grid search? https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a, 2020. Accessed: 2023-05-29.

[38] Google. Google machine learning. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, 2022. Accessed: 2023-05-29.

[39] baeldung. What is a learning curve in machine learning? https://www.baeldung.com/cs/learning-curve-ml, 2023. Accessed: 2023-05-29.

[40] Mary T. Donofrio. Diagnosis and treatment of fetal cardiac disease. https://www.ahajournals.org/doi/10.1161/01.cir.0000437597.44550.5d, 2014. Accessed: 2023-05-29.

[41] Cleveland Clinic. Heart disease. https://my.clevelandclinic.org/health/diseases/24129-heart-disease, 2022. Accessed: 2023-05-29.

[42] Steffen Sjursen. The pros and cons of using jupyter notebooks as your editor for data science work. https://betterprogramming.pub/, 2020. Accessed: 2023-05-09.

[43] Gianpaolo Maso. Classification of fetal heart rate (fhr) pattern subgroups. https://www.researchgate.net/figure/Classification-of-fetal-heart-rate-FHR-pattern-subgroups_tbl2_229082427, 2012. Accessed: 2023-05-27.