

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL
BIG DATA ANALYTICS AND MANAGEMENT HEAD OF THE
DEPARTMENT

HEART FAILURE PREDICTION USING MACHINE-LEARNING
METHODS



MASTER'S THESIS
ESRA MOHAMED OSAMA A. ALGHWIL

ISTANBUL 2024

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL
BIG DATA ANALYTICS AND MANAGEMENT

**HEART FAILURE PREDICTION USING MACHINE-LEARNING
METHODS**



MASTER'S THESIS

THESIS ADVISOR
ASSIST. PROF. GUNET EROGLU

ISTANBUL 2024



T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

MASTER THESIS APPROVAL FORM

Program Name:	Big Data Analytics and Management
Student's Name and Surname:	Esra Mohamed Osama A. Alghwil
Name Of The Thesis:	Heart Failure Prediction Using Machine Learning Methods
Thesis Defense Date:	5 June 2024

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Dr. Yücel Batu SALMAN

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	Institution	Signature
Thesis Advisor's	Assist. Prof. Günet Eroglu	Bahçeşehir University	
Member's	Assist. Prof. Erkut	Bahçeşehir University	
Member's	Prof. Dr. Murat Akbalık	Marmara University	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Esra Mohamed Osama A.Alghwil

Signature:

ABSTRACT

HEART FAILURE PREDICTION USING MACHINE LEARNING METHODS

Esra Mohamed Osama A. Alghwil

Master's Program in Big Data Analytics and Management

Supervisor: Assist. Prof. Gunet Eroglu

May,2024, 120 pages

People nowadays are looking for answers to ailments that threaten human being life and lead to death, and the majority of these illnesses are connected to the heart, which is one of the most crucial organs in the body. As a result, scientists attempted to anticipate these illnesses using previously collected hospitalization data from patients. Therefore, data science has been utilized to evaluate these collected data and anticipate that type of sickness using machine learning tools and procedures. Throughout this paper, analysis of the data and visual analytics were being used on two datasets that included different and various conditions such as high blood pressure, anaemia, and others, as well as health issues such as changes in creatinine phosphokinase or serum creatinine levels, to predict whether such issues and illnesses would cause heart failure or not. This work provides prediction of heart failure using machine learning models and focuses on finding if the length and the variation in the features of the datasets would affect the prediction results or not.

Key Words: Heart Failure; Machine Learning; Prediction; Linear Discriminant Analysis; Data Modeling.

ÖZ

MAKİNE ÖĞRENME YÖNTEMLERİ KULLANILARAK KALP YETMEZLİĞİ TAHMİNİ

Esra Mohamed Osama A. Alghwil

Büyük Veri Analitiği ve Yönetimi Yüksek Lisans Programı

Tez Danışmanı: Assist. Prof. Gunet Eroglu

Mayıs 2024, 120 sayfa

Günümüzde insanlar, insan hayatını tehdit eden ve ölüme yol açan hastalıklara cevap arıyor ve bu hastalıkların büyük çoğunluğu vücudun en önemli organlarından biri olan kalp ile bağlantılı. Sonuç olarak bilim insanları, hastalardan önceden toplanan hastaneye yatma verilerini kullanarak bu hastalıkları tahmin etmeye çalıştı. Bu nedenle, toplanan bu verileri değerlendirmek ve makine öğrenimi araçlarını ve prosedürlerini kullanarak bu tür hastalıkları tahmin etmek için veri biliminden yararlanıldı. Bu makale boyunca, yüksek tansiyon, anemi ve diğerleri gibi farklı ve çeşitli durumların yanı sıra kreatinin fosfokinaz veya serum kreatinin düzeylerindeki değişiklikler gibi sağlık sorunlarını içeren iki veri kümesinde veri analizi ve görsel analiz kullanıldı. Bu tür sorunların ve hastalıkların kalp yetmezliğine neden olup olmayacağını tahmin etmek. Bu çalışma, makine öğrenimi modellerini kullanarak kalp yetmezliğinin tahmin edilmesini sağlıyor ve veri setlerinin uzunluğunun ve özelliklerindeki varyasyonun tahmin sonuçlarını etkileyip etkilemeyeceğini bulmaya odaklanıyor.

Anahtar Kelimeler: Kalp Yetmezliği; Makine Öğrenme; Tahmin; Linear Discriminant Analysis; Veri Modelleme

ACKNOWLEDGEMENTS

My sincere gratitude and high esteem are for Assist. Prof. Gunet Eroglu who was giving me her valuable instruction, advice, and reflections that provided the opportunity to complete this project.

I also care for and appreciate my parents for their kindness, caring. They provide me with the encouragement and comprehension necessary for me to achieve this degree.



TABLE OF CONTENTS

ABSTRACT	iv
ÖZ.....	v
DEDICATION	v
ACKNOWLEDGEMENTS.....	vi
LIST OF TABLES	ix
LIST OF FIGURES	x
Chapter 1	1
Introduction	1
1.1 Theoretical Framework.....	1
1.2 Statement of the Problem	3
1.3 Purpose of the Study.....	4
1.4 Research Questions	5
1.5 Significance of the Study	5
1.6 Definitions.....	5
Chapter 2	8
Literature Review	8
Chapter 3	16
Methodology	16
3.1 Research Design.....	16
3.2 Universe and Participants.....	26
3.3 Data Collection	27
3.3.1 Data Collection Instrument	27
3.3.2 Data Collection Procedures.....	28
3.3.3 Data Analysis Procedures	29
3.3.4 Reliability and Validity	94
3.4 Limitations.....	97

Chapter 4	100
Findings.....	100
Chapter 5	107
Discussion and Conclusion.....	107
5.1 Discussion of Findings for Research Questions	107
5.2 Conclusion.....	110
5.3 Recommendation.....	111
REFERENCES.....	113
Appendix	



LIST OF TABLES

TABLES

Table 1 The Accuracies from The Literature Review.....15



LIST OF FIGURES

FIGURES

Figure 1 Decision Tree Algorithm	18
Figure 2 Naïve Bayes Algorithm Equation.....	19
Figure 3 Random Forest Algorithm	21
Figure 4 The Area Under the Curve	23
Figure 5 Histogram for dataset 1	33
Figure 6 Histogram for the Rest of Dataset 1	33
Figure 7 Histogram for Dataset 2.....	34
Figure 8 Histogram for the rest of Dataset 2.....	34
Figure 9 The First Heat Map for Dataset 1	35
Figure 10 The Second Heat Map for Dataset 1.....	36
Figure 11 The Heat Map for Dataset 2	36
Figure 12 The Pie Chart for the Target of Dataset 1.....	38
Figure 13 The Pie Chart for the Gender of Dataset 1	39
Figure 14 The Bar Chart of Age Vs Outcome	40
Figure 15 The Bar Chart of Atrial Fibrillation Vs Outcome	41
Figure 16 The Bar Chart of Heart Rate Vs Outcome.....	42
Figure 17 The Bar Chart of Respiratory Rate Vs Outcome	43
Figure 18 The Bar Chart of MCV Vs Outcome.....	44
Figure 19 The Bar Chart of RDW Vs Outcome for Males	45
Figure 20 The Bar Chart of RDW Vs Outcome for Females.....	46
Figure 21 The Bar Chart of Leukocyte Vs Outcome	47
Figure 22 The Bar Chart of PT Vs Outcome	48
Figure 23 The Bar Chart of INR Vs Outcome	49
Figure 24 The Bar Chart of CK Vs Outcome for Males.....	51
Figure 25 The Bar Chart of CK Vs Outcome for Females	52
Figure 26 The Bar Chart of Creatinine Vs Outcome for Males.....	53
Figure 27 The Bar Chart of Creatinine Vs Outcome for Females	54
Figure 28 The Bar Chart of Urea Nitrogen Vs Outcome	55
Figure 29 The Bar Chart of Blood Potassium Vs Outcome.....	57
Figure 30 The Bar Chart of Anion Gap Vs Outcome	58

Figure 31 The Bar Chart of Lactic Acid Vs Outcome	59
Figure 32 The Pie Chart of Target for Dataset 2	61
Figure 33 The Pie Chart of Gender for Dataset 2	62
Figure 34 The Distribution Map of Exercise Angina Vs Outcome	63
Figure 35 The Pie Chart for the Types of Chest Pain	64
Figure 36 The Distribution Map of Chest Pain Types Vs Outcome	65
Figure 37 The Distribution Map of Gender Vs Outcome	66
Figure 38 The Distribution Map of Age Vs Outcome	67
Figure 39 The Bar Chart of Age Vs Outcome	68
Figure 40 : The Distribution Map of Fasting Blood Sugar Vs Outcome	69
Figure 41 The Distribution Map of Resting Blood Pressure Vs Outcome	70
Figure 42 The Distribution Map of Resting Blood Pressure Vs Outcome	71
Figure 43 The Distribution Map of Resting ECG Vs Outcome	72
Figure 44 The Distribution Map of the Models Accuracies for Dataset1	82
Figure 45 The Distribution Map of the Models Accuracies for Dataset2	83
Figure 46 The Confusion Matrix of the Naive Bayes Model for Dataset1	84
Figure 47 The Confusion Matrix of the LightGBM Model for Dataset2	85
Figure 48 The Distribution Map of the Cross-validation Accuracies for Dataset1 ...	88
Figure 49 The Distribution Map of the Cross-validation Accuracies for Dataset2 ...	89
Figure 50 The Distribution Map of the Principal Component Analysis Accuracies for Dataset1	91
Figure 51 The Distribution Map of the Principal Component Analysis Accuracies for Dataset2	93
Figure 52 The AUC-ROC of Linear Discriminant Analysis model for Dataset1	96
Figure 53 The Distribution map of Feature Importance for Linear Regression for Dataset1	103
Figure 54 The Distribution map of Feature Importance for Linear Discriminant Analysis for Dataset1	104
Figure 55 The Distribution map of Feature Importance for Cat Boost for Dataset1	105

Chapter 1

Introduction

1.1 Theoretical Framework

Heart is generally the most significant organ in the human being body. This is due to the manner in which it pumps the blood through the human body, in return providing oxygen as well as nutrients to all the body parts. Similarly, it removes the carbon dioxide from a human body that is transported by the blood as returned to the heart. The heart is important for the overall well-being of a person; any cardiac malfunction or accident may lead to severe impacts. In addition to that, all the organs of the body are prone to a range of diseases including the heart. One of these diseases that may affect the heart is heart failure. According to (Mortazavi, et al., 2016), heart disease is the most serious mortality cause in the world. Illnesses such as heart attack, arrhythmias, and heart failure can substantially affect an individual's way of living. As reported by (Malik, Brito, Vaqar, & Chhabra, 2023), heart failure appears when the heart fails to deliver enough blood to the body, resulting in the death of most people affected. The heart failure event may result from several factors, which include different blood pressure problems, diabetes, obesity, anemia, and a vast variety of other heart-dangerous conditions that are prone to affect the heart. Additionally, heart failure might reduce an individual's quality of life through limited ability to do everyday activities or worsen general health. It can also be the reason of the increase in probability of being hospitalized or at worst, dying. Heart failure is normally rectified through lifestyle changes, medication, and surgical improvement or procedures. Therefore is essential to be mindful of heart health by ensuring proper dietary reference, frequent physical activity, measuring primary issues through medical practice.

Heart failure is critical to forecast since it is a significant medical illness that can result in impairment, hospitalization, and possibly death. Early identification and treatment of cardiac failure, as noted in the article by (Pasha, Ramesh, Mohammad, Harshavardhan, & Shabana, 2020), can enhance an individual's standard of life, lower the possibility of difficulties, and raise the odds of survival. Predicting cardiac failure can help reduce the high readmission incidences following heart failure after being hospitalized. It can place a significant cost on individuals and the healthcare system as

a whole. It also aids health practitioners in the identification of persons at higher risk attaining the illness and giving measures and therapies to decrease one's risk. Predicting heart failure might also be helpful for health professionals to use resources more effectively and efficiently. This factor offers an interval for recognizing patients who require more frequent monitoring or more intensive treatment. Cardiogenic failure prediction might raise the likelihood of survival and lower healthcare costs in general. According to (Mortazavi, et al., 2016), prediction models can help detect patients at increased risk of returning to the hospital. By identifying important risk variables, the methods may help the healthcare prescribe specific treatments to individuals to receiving the greatest benefit.

Machine learning can predict cardiac failure by evaluating vast volumes of data drawn from many sources such as biomarkers, clinical risk variables, indicators, and imaging investigations. In (Segar, et al., 2021), the authors highlighted how the models included laboratory, clinical, and biomarker data and outperformed established cardiovascular risk models. The machine learning algorithms identified the most significant determinants of heart failure risk. It needs to be noted that machine learning has multiple advantages in forecasting heart failure. One advantage of employing machine-learning algorithms in building predictive models of heart failure is their ability to analyze large amounts of data from heterogenic sources like medical risk variables, indicators, and imaging tests and expose complex interactions among risk variables. Identifying patients at the highest risk of eventually developing heart failure can lead to better and more accurate predictive risk. They can use machine-learning models, implemented for a particular subgroup, certain demographics, or people identifying as unique ethnic or racial groups, to generate tailored risk assessments. Machine learning models can help identify those at elevated risk of coronary artery disease due to their unique traits and provide tailored preventative measures. The models may also be connected with electronic medical records to determine current risk levels for individuals. It can assist physicians in identifying high-risk individuals and delivering prompt therapies to avert heart failure.

Furthermore, compared with traditional statistical models, as stated in (Segar, et al., 2021), machine learning models are more capable of dealing with missing data on variables and nonlinear interactions among co-variates. Therefore, even when data is insufficient or unavailable, machine learning can still improve the precision of the

probability of heart failure projection. With time, they could possibly be upgraded as more data is available, facilitating evolutionary risk of heart failure projection. In the same way, these models could be kept catered to the test, meaning current and useful. Machine learning, in general, has made a significant contribution to increasing the likelihood of predicting heart failure with accuracy. It aggregates data from multiple sources and finds complex links between risk factors. It can be used to identify patients at risk and can thereby be included in preventive strategies and, more generally, heart failure management.

1.2 Statement of the Problem

Globally, heart failure is considered one of the most severe problems. As previously said, it is a critical medical illness in which the heart fails to circulate sufficient blood to satisfy the body's demands. It can cause various symptoms, including difficulty breathing, weariness, leg and ankle swelling, and trouble exercising. It was discussed how cardiac disorders, particularly heart failure, are a primary cause of mortality and disability globally in (Vaduganathan, Mensah, Turco, Fuster, & Roth, 2022). The research also includes heart failure risk factors, such as - Cardiovascular disease - Sleep apnea - Smoking - Heart disease in the family - Diabetes - High blood pressure - Abuse of alcohol - Obesity is a problem. - Cardiovascular disease - Cigarette smoking - Age (risk rises with age) - Certain drugs or medical problems, like chemotherapy or thyroid diseases, can harm the heart. The fact that some factors are unchangeable, such as age and family history, should not be overlooked. However, lifestyle modifications and medical therapy may address other risk factors. (Vaduganathan, Mensah, Turco, Fuster, & Roth, 2022), underlined that heart failure should be addressed due to its potential consequences, such as renal damage, damage to the liver, and even mortality. As a result, predicting heart failure would help individuals avoid becoming ill, which is how machine learning is used to assist in the healthcare unit. As previously discussed, machine learning has numerous approaches that may be utilized for training data and using the trained data for prediction. In this work, these methods were employed on two datasets for predicting heart failure.

Although there is some work has been done on predicting heart failure using machine learning, there has yet to be a comparison between two datasets, one that has a large number of patients and the other does not, so the outcome of the comparison remains

determined. In this study, two datasets were taken and compared to determine which would give better results than the other to predict heart failure. Furthermore, it is worth noting that while these datasets have different variables, they were trained and subjected to the same machine-learning method in order to achieve good results. Moreover, the comparison also aims to determine which dataset would provide better accuracy. In addition, it aims to determine whether the length of the dataset affects the accuracy.

1.3 Purpose of the Study

This work aims to provide an accurate prediction of heart failure by employing machine-learning methods. The sense behind this is that it is crucial to build a structure that can produce rules or identify data using machine-learning approaches, as well as to boost knowledge of the root causes and risk factors related to the condition so that healthcare professionals can make timely diagnoses of heart failure by using this information. The purpose is to avoid the risky scenario of a rising number of heart failure individuals and to enhance the odds of recognizing heart disease in people. According to (Adler, et al., 2019), past heart failure prediction techniques were mainly produced using statistical and analytical approaches, which failed to consider complex correlations that include prognostic information. Machine learning, on the other hand, has long been utilized in other domains, including high-energy physics. It could be used to develop a model, which more precisely forecasts the likelihood of death among individuals suffering from heart failure. As a result, machine-learning models may be employed to analyze medical records and forecast the possibility of heart disease in individuals, assisting doctors and heart surgeons with accurate diagnoses and delivering prompt treatment. In predicting heart failure, healthcare practitioners may identify individuals at risk of acquiring the illness and take preventative actions to limit its occurrence. Machine learning approaches may also be used to create prediction models that can assist healthcare practitioners in the early detection of heart failure and in selecting the most effective therapies. Furthermore, machine-learning techniques may be utilized to uncover patterns of patient information and potential therapies that could assist in minimizing the burden of heart failure. The ultimate objective of utilizing machine-learning approaches to predict heart failure is to enhance the level of treatment for people struggling with the ailment.

1.4 Research Questions

How to predict heart failure using machine learning?

Which method gives better accuracy?

Which dataset gives better results?

Does the dataset length affect the results?

Does the dataset variables affect accuracy?

1.5 Significance of the Study

This article has helped to learn the potential application of machine learning when predicting heart failures. To ensure the highest accuracy possible while predicting heart failures with the results of studying two datasets, this study has conducted several machine-learning methods for two datasets. By applying machine learning models and techniques to the process of diagnosing heart failure, healthcare professionals would be able to help patients better. Moreover, within this article, it has been investigated whether the use of the machine learning method provides the proper predictions for the two datasets that were studied. It is essential to learn whether the length or the features included in both datasets would affect the prediction's accuracy.

1.6 Definitions

Machine Learning: is the area of computing, which has been growing in the last few years, and it aims to use the intelligence of humans by learning from their surroundings. Moreover, it is known as the workhorse for big data.

Decision tree: An acyclic-directed graph that is constructed by recursively partitioning a dataset from the top down. The method is known for its power for regression and classification, and it is a well-known technique in machine learning.

K- Nearest Neighbor (KNN): It is a model in machine learning that is considered a simple and effective model that is used for prediction and classification. A primary use of the algorithm is to determine which pattern is closest to the target pattern in calculating label information. By using KNN, most K-nearest patterns in data space can be classified as classes and thus categorized.

Area Under the Curve (AUC): It is a metric that can be considered to measure how the value of a variable changes over time and to assess the overall pattern or total

amount of change. Typically, this method is used when a response variable must be obtained many times during the study, such as salivary cortisol.

Support Vector Machines (SVM): They are a tool for predicting regression and classification that employs machine-learning concepts to enhance predicted accuracy while minimizing over-fitting to the data. SVMs are considered models that utilize the hypothesis space of linear functions modeled in high-dimensional feature spaces using a learning algorithm derived from optimization theory that implements a statistical learning bias.

Artificial Neural Network (ANN): It is a computer model inspired by the shape and function of biological neural networks. Usually, they are composed of several nodes that are connected, called processing elements (PEs), which mimic the actions of neurons. As a result of the design of ANNs, they can be used for learning and processing data generated from inputs and for predicting patterns, recognizing patterns, and classifying them.

Random Forest: It is a form of group approach that is based on trees and is used for sorting and analyzing data. They are created to compete with boosting. In classification and regression applications, Random Forests could be used to handle categorical and continuous predictor variables. Random Forests have many attractive features, such as handling high-dimensional issues, quick training and prediction, and assessment of generalization errors.

Naive Bayes: A learning method that is based on Bayesian learning principles and relies on the premise that the characteristics are conditionally independent concerning the class known. As a result of its computing economy and competitive classification accuracy, it is often used in practice because of its efficiency in computing.

Logistic Regression: A methodology of building a statistical model for a dataset in which the dependent variable is binary. The possibility of occurrence of specific values of the dependent variable is expressed as a logistic regression equation when the values are set by the independent variable.

Linear Discriminant Analysis: It is a model for categorizing data, essentially developing a weighted average of some variables to minimize the dimensions of those variables and utilize as many of them as feasible to distinguish categories.

Cross Validation: A cross-validation technique is used to determine the prediction error of statistical prediction functions and to evaluate statistical models. In this method, the dataset is applied to separation into two sets, a training and a testing set, which enables recurrent model testing across multiple divisions of data.

Principle Component Analysis (PCA): It is considered as a statistical method for extracting and visualizing the relationship between data sets by transforming the variables into a few independent variables or components. PCA is used mainly in reducing the factors in the composite or multiple correlation coefficient and its estimation clustering.

CATBoost: It is a general-purpose cat boost gradient boosting implementation used to increase gradients on decision trees for classification, regression, or ranking. This implementation is concerned with working on easier projects. The CatBoost algorithm performs well on large and complex datasets with categorical characteristics using gradient-based optimization, ordered boosting, and random permutations.

Extra Trees: It is a technique used by machine learning, which combines the output of multiple decision trees in order to provide predictions based on their trained input.

Adaptive Boost: The adaptive boosting technique, or AdaBoost, is a supervised learning method that classifies data by combining a number of weak or base learners into a strong model. To make AdaBoost work, the training dataset's instances are weighted based on how well prior classifications performed.

Extreme Gradient Boosting: XGBoost is a free software library that performs optimal distributed gradient boosting machine learning methods. With its parallel tree boosting capabilities, it is one of the most powerful machine learning packages for classification, ranking, and regression problems.

LightGBM: LightGBM is a library for gradient boosting that is accessible, distributed, and powerful. The design of the LightGBM method puts a high priority on accuracy, scalability, and efficiency. Essentially, the model is based on decision trees, which are intended to reduce memory utilization and enhance model performance.

Chapter 2

Literature Review

Heart failure is a serious disorder with incidence and expense. Several types of research have focused on predicting coronary artery diseases like heart failure using machine-learning models. Naïve Bayes, k-nearest neighbors, Random Forest, and others are the most widely employed approaches. These approaches are often used on extensive datasets which contain demographic and clinical data, including blood pressure. In addition, they include behavioral aspects like smoking and food. Scientists want to improve the accuracy of heart failure prediction models by adjusting variables and integrating more factors, which will result in improved preventative therapy for at-risk patients.

According to (Adler, et al., 2019) study, they applied a machine learning model called boosted decision trees (BDTs) to estimate mortality rates among heart failure patients. Based on a sample of patient records, the method was developed to determine a combination of eight characteristics that demonstrated excellent discrimination between high-risk and low-risk individuals. External validation in two different heart failure groups yielded AUCs of 0.84 and 0.81, higher than those obtained using the two current risk estimates for these patients. As a result, machine learning computed the mortality rates in heart failure individuals with significant accuracy.

Moreover, in (Javeed, et al., 2022) article, they used other machine learning techniques such as support vector machines (SVM), K-nearest neighbors (KNN), along with neural networks fuzzy modeling and others, which are utilized for automated detection of cardiac disease. The study considers the importance of data in ML-based, particularly when deep learning models are utilized, and the requirement for huge datasets for higher classification performance. The researchers highlighted the difficulties in employing deep learning technology for heart disease diagnosis, such as the need for massive data for model training and the temporal complexity of ML models.

In the (Jindal, Agrawal, Khera, Jain, & Nagrath, 2021) study, they used similar machine learning methods such as Random Forest Classifiers, K nearest neighbors (KNN), and Logistic Regression. The researchers combined Logistic Regression and K nearest neighbors models to clean the data, and this combination gave an accuracy

of 87.5%. However, when they used each model, the KNN gave them the highest accuracy (Table I). Likewise, in (Srivastava & Choubey, 2020) article, they suggested a technique based on several algorithms, concluding that K-Nearest Neighbors give the highest accuracy (Table I).

Furthermore, (Pasha, Ramesh, Mohmmad, Harshavardhan, & Shabana, 2020) examined a Kaggle dataset that included variables associated with heart diseases like age, blood pressure, gender, cholesterol, and others. They examined the degrees of accuracy of several machine-learning algorithms, including Decision Trees (DT), Support Vector Machines (SVM), Decision Trees (DT), Artificial Neural Networks (ANN) and K-Nearest Neighbor (KNN). The data showed the ANN model's most remarkable accuracy (Table I). However, when (Sahoo & Jeripothula, 2021) investigated the efficacy and accuracy of various algorithms for categorizing heart diseases based on their risk factors and symptoms. The result that they came up with showed that SVM was the most accurate method (Table I).

Moreover, in (Chauhan, 2018) article, he described the application of machine learning models for heart disease risk prediction based on 14 independent factors. The study covers data analysis, methodology, and classification algorithm implementation. The logistic regression model has the highest accuracy, which makes it the most efficient method for the study (Table I). However, it is essential to note that SVM and Random Forest provide accuracy similar to Linear Regression (Table I). In addition, (Rindhe, Ahire, Patil, Gagare, & Darade, 2021) investigated the importance of machine learning algorithms for predicting cardiovascular illnesses and selecting individuals for preventative therapy at an early stage. They discovered that the most accurate accuracy attained in the research utilizing Support Vector and Neural Network classifiers, respectively (Table I).

In (Lakshmanarao, Swathi, & Sri Sai Sundareswar, 2019) paper, they used a dataset with 15 patient-specific variables to determine if a person has a 10-year risk of future heart attack or stroke. The study utilized the following methods: random oversampling, synthetic minority oversampling (SMOTE), and adaptive synthetic sampling. Using methods based on machine learning, it reached 99% accuracy in detecting heart illness. The findings demonstrate that SVM performs best for random oversampling, whereas Random Forest and Extratree Classifiers perform best for SMOTE and adaptive synthetic sampling.

In relation, (Tripoliti, Papadopoulos, Karanasiou, Naka, & Fotiadis, 2017) examined in their article various machine learning models used in the management of heart failure, such as prediction of hospital readmission and adverse events, as well as diagnosis and classification of heart failure subtypes. They tested numerous approaches and procedures for detecting, classifying, estimating severity, and predicting adverse events in heart failure patients.

Other method was used by (Bharti, et al., 2021), which describes how to apply both machine learning and deep learning algorithms for cardiovascular disease prediction using the UCI Machine Learning Heart Disease dataset. They illustrated the machine learning and deep learning techniques used to predict heart disease, such as K-nearest neighbors, support vector machines, random forests, logistic regression, decision trees, and deep learning algorithms. The article goes through numerous assessment measures for deep learning and machine learning techniques, such as the confusion matrix, recall accuracy score, sensitivity, precision, accuracy score, and F1 score. The findings were impressive, with the accuracies that were given using deep learning and KNN (Table I).

In addition, in (Haq, Li, Memon, Nazir, & Sun, 2018) article, they examined the use of three alternative feature selection methods for the detection of heart disease: LASSO, MRMR (Maximum Relevance — Minimal Redundancy), and Relief. The findings of the trials demonstrate that classifiers performed very well on six crucial characteristics. The logistic regression performed well, with an 84% classification accuracy, an 85% specificity, and a 77% sensitivity. Logistic regression and SVM (RBF) (Support Vector Machines with Radial Basis Function kernel) have higher ROC AUC (area under the receiver operating characteristic curve) values than other classifiers, at 88% and 87%, respectively.

El-Hasnony, Elzeki, Alshehri, and Salem (2022) explained in their article the application of active learning (AL) models for prediction of heart disease using machine learning (ML). The tests were meant to examine the accuracy of AL models and F-score to improve the rate of recognition. The grid search approach was used to optimize hyperparameters in the tests. The findings show that the Adaptive selection approach may produce a greater rate of accuracy for a learned model with a generalization idea than one with a memorizing concept, which can be obtained using

a traditional ML model (s). However, the results could have been more accurate. The recall gave an accuracy of 65%.

As well, in (Princy, Parthasarathy, Hency Jose, Raj Lakshminarayanan, & Jeganathan, 2020) study, the Decision Tree algorithm scored the greatest prediction accuracy when compared to the other classifiers (Table I). This percentage was not as accurate as it could be due to the dataset's dimension, which influenced algorithm performance.

On the other hand, (Chandra , Shue Nee, Zhi Min, & Xin Ying, 2019) study demonstrated that employing the random forest algorithm would be a better alternative for more accuracy. With an accuracy of 90-95%, it is a functional classification system for accurate Heart Disease Prediction. The study also implies combining non-modifiable and modifiable risk variables may result in better prediction of heart disease.

Absar, et al. (2022) used in their study the pre-filtering and trimming of data to fit the model better and used a variety of hyper-parameters and training settings to train the model more precisely. Because it employed two datasets, demonstrating that the k-nearest neighbor and decision tree classifiers for the first dataset (Cleveland) and the random forest classifier for the second dataset produced the maximum accuracy (CHSLB) (Table I).

On the other hand, (Olsen, Mentz, Anstrom, Page, & Patel, 2020) used and developed a convolutional neural network (CNN) algorithm that instantly calculated left ventricular ejection fraction (LVEF) and longitudinal strain (LS) utilizing electrocardiography data.

Angraal, et al. (2020) discovered in their study that the Random Forest (RF) classifier performed best, with a mean C-statistic of 0.72 in predicting death and 0.76 in predicting HF hospitalization. The research also found that general health data, as measured by Kansas City Cardiomyopathy Questionnaire (KCCQ) scores, was one of the most powerful predictors, including death and HF hospitalization.

Wang (2021) examined in his article the efficacy of 18 machine-learning models depending on 12 clinical variables, and the models were tested using the F1 score and accuracy as assessment measures. The researchers discovered that z-score

normalization is superior to min-max normalization and that utilizing SMOTE for class imbalance is beneficial.

Moreover, (Nikam, Bhandari, Mhaske, & Mantri, 2020) developed machine-learning algorithms for predicting cardiovascular disease using characteristics, emphasizing the influence of body mass index (BMI) on the prediction. Examining the models reveals that BMI is critical in boosting the model's accuracy. Consequently, the decision tree classifier was shown to be the most efficient (Table I).

Lutimath, Chethan, and Pol (2019) in their work they investigated the use of machine learning methods for predicting cardiac disease, including Naive Bayes classifiers and radial kernel support vector machines. According to the article, different strategies are compared for their accuracy and it is discovered that the radial kernel SVM provides a higher degree of accuracy than the Naive Bayes classification.

Furthermore, (Ansari, Alankar, & Kaur, 2020), suggested a machine learning-based methodology for accurately predicting cardiac disease. The researchers examined the performance of several methods and suggested a modified technique that combines logistic regression with the analysis of principal components. The model that gave the highest accuracy was logistic regression using PCA (Table I).

In the work, (Nazrul Anuar, et al., 2020) , they used supervised machine learning approaches to construct an automated classification model for predicting CVD risk from ECG data along with additional clinical. The model that performed the best in terms of specificity, sensitivity, and accuracy 90% was a two-layer feed-forward and back-propagation neural. Because it optimized the amount of hidden layers, learning method, and transfer function employed, ANN accuracy.

Chicco and Jurman (2020) in their study they examined ways in which machine learning algorithms could be used to predict heart failure survival using only two features, i.e. the ejection fraction and serum creatinine. It was found that ejection fraction and serum creatinine are the two most relevant characteristics for predicting, and it was found that Random Forests was the most reliable classifiers (Table 1).

In the research, (Barhoom, Almasri, Abu-Nasser, & Abu-Naser, 2022), explained how they integrated deep learning to forecast cardiac. They employed under and over-sampling approaches to preprocess and balance the data. Deep learning achieved the highest accuracy with under and over-sampling, with 78.77% and 92.35%

accuracy, respectively. Random Forest had the second-highest over-sampling, at 90.21% accuracy.

Saboor, et al. (2022) they also employed several machine-learning algorithms in their work to predict cardiac illnesses, but with hyperparameter tuning, which increased the algorithms' accuracy. It was found that hyperparameter tuning enabled the SVM classifier to achieve the highest accuracy of all the classifiers due to the experimental findings. The accuracy obtained from the SVM classifier was 94% precision, F-measure, recall, and 96.72% accuracy.

Additionally, (Choudhary & Narayan , 2020), examined numerous characteristics and used decision tree and Ada-Boost algorithms to determine whether patients had cardiac disease. Although the decision tree model overestimates the data, the Ada-Boost method outperforms it (Table I).

In the research, (Hamdaoui, Boujraf, Chaoui, & Maaroufi, 2020), They developed a clinical assistance system for detecting cardiac disease utilizing machine learning techniques, including Support Vector Machine, Decision Tree, Naive Bayes, and others. Applying cross-validation with train-test split approaches, the researchers discovered that Naive Bayes beat the other algorithms with an accuracy of 82.17% and 84.28%, respectively, on the UCI Cleveland heart disease dataset.

Moreover, in (Lorenzoni, et al., 2019) they examined the performance of eight machine learning approaches' performance in predicting heart failure patients' hospitalization. The researchers found that the Generalized linear model (GLM) performed the most accurate in terms of hospitalization prediction, with an average accuracy of 81.2%, a positive predictive value of 87.5%, and a negative predictive value of 75%, respectively, according to the researchers.

The aim of (Krishnani, Kumari, Dewangan, Singh, & Naik, 2019) study is to determine the significance of traits in categorizing chronic illnesses, notably coronary heart disease. The dataset was preprocessed and normalized to produce comprehensive, consistent, and interpretable data. The comparison of machine learning algorithms revealed that the Random Forest approach had the most accurate performance measure among all classifiers (Table I).

As a final point, it is imperative to mention that some of the studies showed a low percentage of accuracy when it came to the results. The reason for this may be due

to a number of factors that affect the accuracy of the classifiers. A lack of data in the dataset is one of the significant factors that could affect the accuracy of the results. In addition, the quality and choice of features are also critical to the model, as well as the complexity of the dataset. As a result, this study took that kind of problem into account.



TABLE 1

The Accuracies from The Literature Review

Sr. No.	Author	Year	Accuracy
1	Jindal ¹ , H., Agrawal ¹ , S., Khera ¹ , R., Jain ² , R., & Nagrath ² , P.	2021	KNN gave the highest accuracy, 88.52%
2	Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P.	2021	Deep learning and KNN gave the highest accuracies, 94.2% and 83.29% respectively
3	Pasha, S. N., Ramesh, D., Mohmmad, S., Harshavardhan, A., & S.	2020	ANN gave the best accuracy with rate of 85.24%
4	Princy, R. J. P., Parthasarathy, S., Hency Jose, P. S., Raj Lakshminarayanan, A., & Jeganathan, S.	2020	Decision Tree gave the highest accuracy of 73%
5	Absar, N., Das, E. K., Shoma, S. N., Khandaker, M. U., Miraz, M. H., I. Faruque, M. R., Tamam, N., Sulieman, A., & Pathan, R. K.	2022	KNN gave 100% accuracy for Cleveland dataset, and RF gave 99.025% for CHSLB dataset
6	Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S.	2020	Decision Tree gave 73.12% accuracy
7	Sahoo, P. K., & Jeripothula, P	2021	SVM gave an accuracy of 85.2%
8	Srivastava, K., and Choubey, D. K.	2020	KNN gave an accuracy of 87%
9	Ansari, M. F., Alankar, B., & Kaur, H.	2020	Logistic Regression gave the highest accuracy using PCA, 86%
10	Chicco, D., and Jurman, G.	2020	RF gave the best accuracy, 74%
11	Chauhan, Y. J.	2018	LR gave the highest accuracy of 89% and similar accuracies came from SVM and RF , with an accuracy of 88% and 87%, respectively
12	Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M.	2021	Support Vector and Neural Network gave 84% and 83.5% respectively
13	Choudhary, G., and Narayan Singh, S.	2020	Ada Boost method gave an accuracy of 85%
14	Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S.	2019	Random Forest gave the highest accuracy, which was 96.71%

Chapter 3

Methodology

3.1 Research Design

To do this study, two different datasets were selected in order to understand whether the different features and different lengths of the datasets would have an impact on the results of prediction. For this reason, these datasets were collected so that they have differences not only in their features but also in their lengths. The rationale for selecting these datasets is that the features of the data are related to the independent variables used for the prediction of heart failure. There were several variables identified as important indicators for predicting heart failure such as age, gender, blood pressure, cholesterol accumulation, chest pain, fasting blood sugar, and other features. Moreover, the length of datasets was established in the selection process following the availability of data and in consideration of whether the length would influence the predictability accuracy, hence the need for an extensively longer set and shorter one.

Once the datasets were collected, they were exposed to the pre-processing subsystem through the identification and filling in any existing missing values. Like every learning algorithms, data pre-processing is a necessary exercise due to the importance of data quality and reliability. As such, the pre-processing subsystem focuses on the identification and subsequent treatment of missing values as a way of ensuring the completeness of datasets and their reliability, hence accuracy and predictability of machine learning results. As well, to assess the datasets, each one was visualized to identify the features most closely connected to the target variable. When applying the test and train process to the models, it is crucial to understand the features and their relationship to heart failure in order to enhance model accuracy. The test and train process incorporates heart failure-related features and excludes those unrelated to heart failure. By eliminating all non-heart failure related features from the testing and training process, the model can maintain a more focused and specific concentration on the task at hand. Reducing noise and irrelevant information in the model allows it to better capture the patterns and relationships that are truly indicative of heart failure, leading to improved model accuracy. Having a good accuracy numbers are very important for healthcare professionals. It helps them identify patients who are prone to heart failure. This allows the condition to be better managed and treated. The

improved specificity also makes it easier to identify potential treatments for heart failure. Healthcare providers can use the increased accuracy to provide more effective care. Patients with a higher risk of heart failure can be closely monitored. Early intervention and preventative measures can be taken. Additionally, potential treatments can be explored and tested. This leads to improved outcomes and better quality of life for patients. Better accuracy rates play a crucial.

After the visualization element was finished, the datasets had been equipped to be carried out to the machine learning techniques. At first, the splitting approach become implemented to break up the dataset into subsets. In machine learning, setting apart the dataset contain the motion of partitioning the dataset into subsets, which is utilized to train and test the dataset. This approach is used to estimate the generalization performance and to modify the hyper-parameter of the model. Data splitting is likewise required to degree the accuracy and mistakes of the model, as well as to discern out the best parameters for machine learning. This can be done by using splitting the data into testing and training out subsets. A training set is utilized to fit the model, and a test set used to evaluate the overall performance of the model on unknown data.

Following the division of the dataset into training and testing subsets, the decision tree and other method were used in the machine learning process. The decision tree method is a common classification and regression technique. Decision tree classifiers are well-known for their improved perspective of performance results, strong accuracy, better tree pruning approaches, and, optimized splitting parameters. A decision tree is a tree-based approach that each path from the root is defined by a data-separating series until a Boolean result is gained at the leaf node. It is a logical structure of knowledge interactions that includes nodes and links. When classification is performed using relations, decision tree nodes indicate purposes. Every node in a decision tree indicates an element or characteristic, and the different branches that emerge from the node indicate the many values that the feature might take until a conclusion or outcome is achieved at the leaf nodes. The leaf nodes in a decision tree reflect the ultimate outcomes or judgments obtained after the data has been repeatedly divided depending on the variables at each node, as shown in the figure (1).

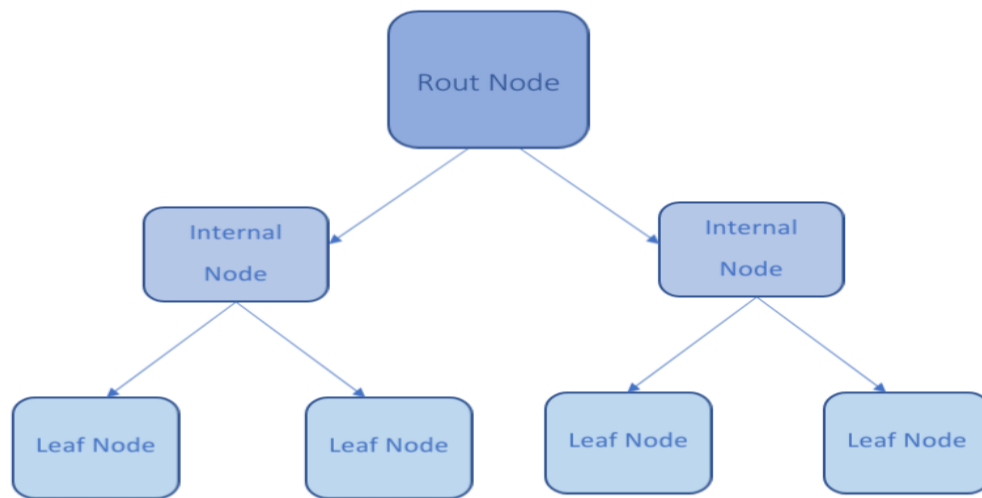


Figure 1 Decision Tree Algorithm

The Naive Bayes machine learning technique was applied as well. Naïve Bayes is a classifier algorithm which is based on Bayes' rule including the premise that features are conditionally independent within a class. It computes the posterior probability of each class given an item, which may then be utilized for classification or decision support applications, as seen in figure (2). Naïve Bayes is computationally effective, enables incremental learning, and is resistant to missing values and noise. It works by computing chances from training data and modifying them when more data becomes available, which allows direct prediction of posterior probability. The method is built on a probabilistic structure that gently decreases in effectiveness in the presence of missing value of attributes. Naïve Bayes is capable of classifying new information that it has not seen previously. This makes it perfect for online learning since newly collected information may be entered without recalculating the probability. Furthermore, Naive Bayes can handle massive amounts of data, which renders it ideal for large-scale applications.

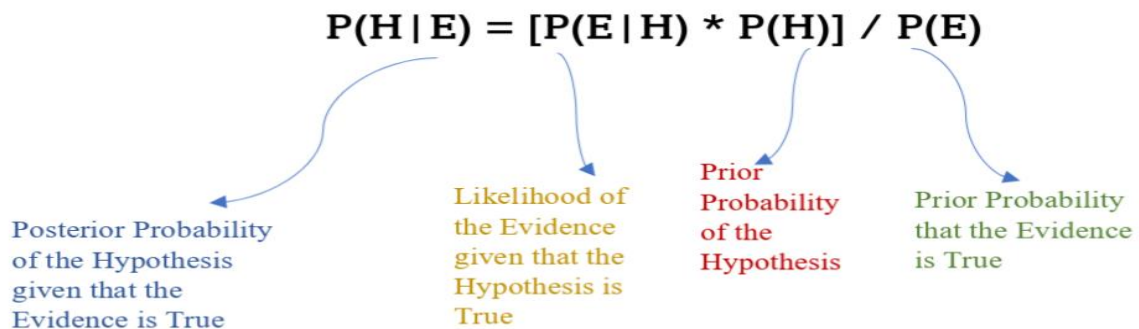


Figure 2 Naïve Bayes Algorithm Equation

KNearest Neighbor (KNN), is another algorithm used, which is a well-known classification model for data mining and machine learning, since it's easy to understand and takes little time to run. KNN categorizes data points according to the vast majority class of their k-nearest neighbors. This technique uses metrics which include Euclidean distance to calculate the gap between records factors and classify them in keeping with their nearest neighbors. The Knn algorithm, based on the majority of its k most similar training data points in the feature space, predicts which test point is to be labeled. In this approach, the parameter k is of particular relevance as it determines how many neighbors are taken into account for a data point classification. Only two variables are needed: k and the similarity or similarity score, to perform this function. These dimensions determine how the classifier will interact with the classification of the training set. The number of neighbors to consider when defining a data point is determined by the K parameter of the method KNearest Neighbors. The value of K is important in this approach because it controls the number of nearest neighbors at each classification step.

Logistic regression (LR), one of the important methods, was used as well for this project. LR is a statistical technique for analyzing dichotomous or binary outcomes with two levels of contradiction. It supports the inclusion of continuous or cluster predictors, as well as correcting for multiple factors, which contributes to the analysis of observational data where the potential for bias needs to be reduced. Logistic regression depends on the odds of a two-level result, in which the odds of an event occurring are calculated as the ratio of the event's likelihood to its probability of not occurring. Logistic regression does not have a formula for estimates, requiring

incremental improvement until stability is achieved, which makes it more difficult to understand by most researchers than linear regression. Due to its ability to cope with the unpredictable variability of these data, logistic regression is preferred over Linear Regression for Binary Results.

There is also the Random Forest method, which was used. Random Forest is defined as an ensemble predictor, which improves accuracy by combining predictions from several occurrences. It is generated utilizing samples from bootstraps of training data and random selection of features in tree induction, figure (3). Random Forest is economical and can be trained quicker than just one decision tree, which makes it ideal for dealing with high-dimensional data. It is an effective technique for compound categorization, capable of producing high prediction accuracy. Random Forest functions by lowering variance with little influence on bias, and thus is useful in situations where variation is a major component of error rate. The approach operates by disrupting the training data and modifying the optimization process, resulting in a decrease in error rate. The bias-variance decomposition of the error rate explains why Random Forest is good in reducing mistake rates. Random Forest chooses essential characteristics but does not generate a detailed model. The alteration in prediction accuracy or margins caused by changing a descriptor can be used to calculate its significance. Random Forest also provides an inherent proximity metric for grouping molecules according to relevant characteristics.

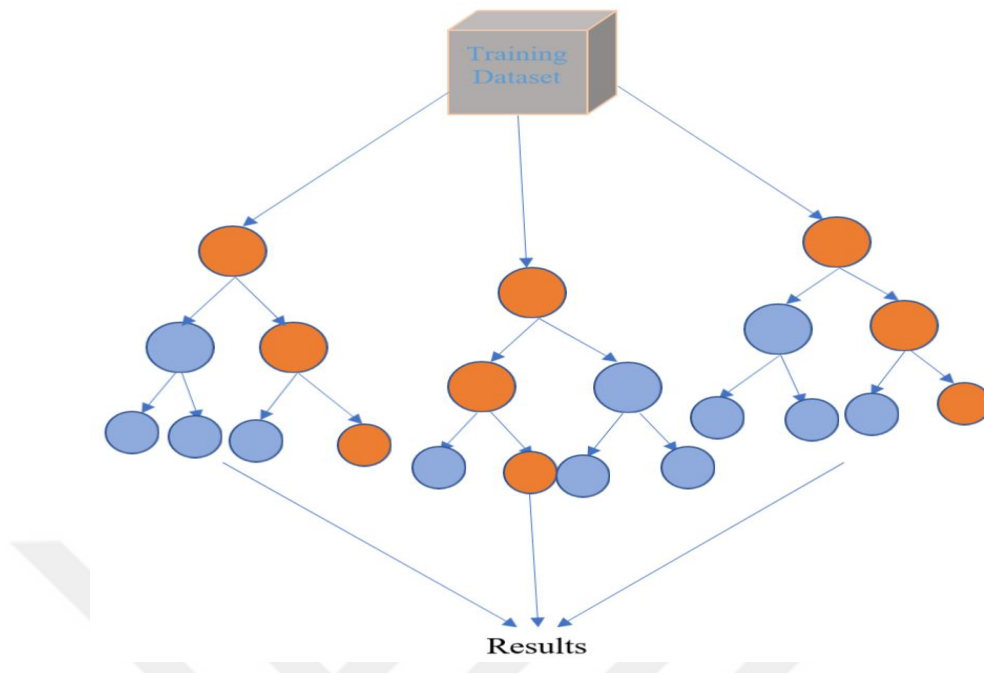


Figure 3 Random Forest Algorithm

Linear Discriminant Analysis (LDA) is also an algorithm used in this study for data classification and dimensionality reduction. Its purpose is to optimize the ratio of between-class variation to within-class variation in a dataset, resulting in maximum separability. LDA can manage scenarios while within-class frequencies are uneven and is extensively utilized in classification tasks, including speech detection. LDA makes the data points rotate along highly discriminating eigenvectors axes acting as boundaries for general classifications. There are two ways in which LDA can be divided as class-dependent and class-independent by deciding according to the number of classifications and the purpose of the data set. Class-dependent LDA increases the amount of between-class variation to within-class variation, which can make the discrimination better. On the other hand, class-independent LDA increases the amount of total possible variation to within- class variation.

This project used yet another method and is used by computers that give labels to the stuff based on an example. The technique is a support vector machine, which forms a hyperplane, learning from labeled data and assigning the labels of objects by examining labeled data. Support vector machines work by enlarging the margin between the class's hyperplane and the nearby data points to divide the class in a high-dimensional linear space as possible. The Kernel functions is put to use by SVM's,

which allows separation of the data points to be done more easily by projecting from a proportional high-dimension tally. It is essential to note that generalization to previously unseen data points may be enhanced if maximizing the distance between the closest data point of each class and the hyperplane. Therefore, as SVMs can define kernels with essentially non-vector inputs and allow various forms of non-vector data to generate diverse kernels, these machines are a helpful tool for analyzing non-vector data in numerous biological applications.

An algorithm was also used in this project, Artificial Neural Network, which is a C++ library specifically designed to search for nearest neighbors in spaces of various dimensions, both exact and approximate. ANN supports the storage of data files in main memory and assumes that points in a given dimensional space are represented as vectors of real numbers or integers. As part of the algorithm, Minkowski metrics are used to determine the distance between points, including Euclidean, Manhattan, and maximum distances. With ANN, the closest (or often k nearest) points are effectively exported to a query point by preprocessing the input points into a data structure. Nearest neighbor queries supported by ANN are both exact and approximate, with approximate queries offering much faster execution times than exact queries.

Additionally, the CatBoost technique was employed. This gradient-boosting toolkit performs better than previous implementations on a variety of datasets. It uses ordered boosting and a new method for processing categorical information to overcome prediction shifts resulting from target leakage. CatBoost relies on decision trees in two modes: Plain and Ordered boosting. Even though the Plain is faster, the Ordered mode is more advantageous for small datasets.

Moreover, CatBoost introduces subsampling and feature combinations to improve the speed of implementation. In addition, to overcome the gradient boosting prediction shift problem, CatBoost uses an ordered boosting method. Furthermore, CatBoost ensures that the model is trained based on different data points. This ensures that the model is trained in a way that it does not use the target of the same data points to prevent bias and ensure generalization by using multiple permutations of the same training data at any step.

Another method used is the AUC, or Area Under the Curve, which is often used as a shorthand for comparing classifiers based on how good they are. It is determined by calculating the enclosed area under the Receiver Operating Characteristic curve, as

seen in Figure (4). It compares the True Positive Rate to the False Positive Rate. It assesses the efficacy of the binary classification model as a whole. The TPR and FPR are all between 0 and 1 in value, which means that the area will similarly always be between 0 and 1. If the AUC value is increased, the model's performance will be improved. The main goal is to enhance the capacity of the area to generate the smallest FPR and the largest TPR for the given threshold. AUC is employed to evaluate classifiers, but this parameter cannot determine which classifier is better for a given target problem.

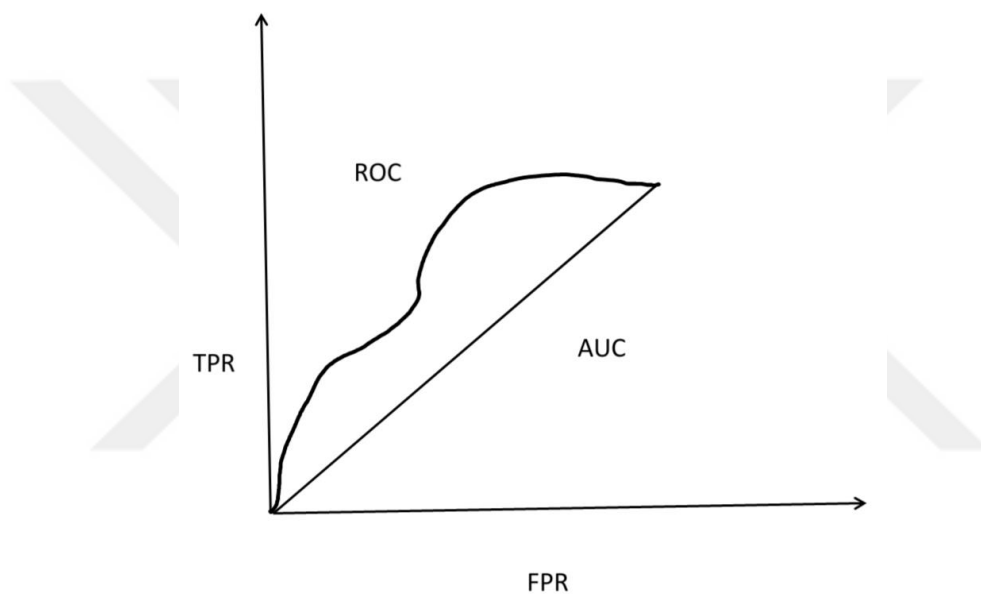


Figure 4 The Area Under the Curve

Likewise, AdaBoost was employed. The AdaBoost method is a boosting strategy that speeds up machine learning algorithms by merging weak classifiers into an effective classification with proper weights. It falls under the category of boosting algorithms and is based on the sequential development of successive classification models. The technique updates the weights of misclassified samples in an iterative manner that ensures that they are emphasized in the next round in order to be correctly classified in the subsequent rounds. The initial selections of poor classifiers significantly affect the quality of classification at the end. AdaBoost works by generating basic classifiers throughout the learning phase while generating

classifications using a voting method. The technique seeks to improve classification accuracy by merging various classifier classes and forecasting algorithms.

For the supervised learning problems with one target and numerical input variables, the non-parametric method, so-called Extra-Trees (ET), was utilized. It means that the whole original training sample is used to create many trees and form the ensemble model. Besides, the cut-point and attribute choice will be selected randomly at each node. Moreover, to receive the prediction at the end, the extra trees provide an arithmetic average in regression cases and consider majority votes in the classification matters. By employing explicit randomized and ensemble averaging, the technique seeks to minimize bias while reducing variance utilizing the whole original learning sample. The Extra-Trees splitting technique for numerical attributes entails choosing a random number of attributes for each node and the smallest sample size for dividing a node.

XGBoost is another technique implemented in the project, and it is the gradient boosting idea; XGBoost adds weak learners in parallel using multi-threading and makes solid CPU usage since it does not use all CPU cores like gradient boosting. Indeed, some of the critical capabilities of XGBoost are continuous training, allowing for improvement of a model that has previously been fit on fresh data, an automated way to handle missing data, and the block structure that allows the parallel construction of the tree. Xgboost is used in applications comprising classification, regression, and predictive modeling and has been commonly successful performing and very effective on structured or tabular datasets.

In comparison to conventional gradient boosting methods, XGBoost builds single trees using multiple cores and arranges data to reduce lookup times on training and prediction. XGBoost has been less informative regarding requirements concerned with the `query_type` parameter.

Another technique used is LightGBM, an open-source, high-performance gradient-boosting framework. LightGBM is created mainly to use relevance, scalability, and preciseness memory. It is based on a decision tree algorithm, but it is intended to decrease memory consumption and improve model performance. LightGBM makes trees with histograms quickly. LightGBM works better than other gradient boosting frameworks due to some mechanisms it employs, as well as enhancements such as leaf-wise tree growth and efficient data storage formats.

LightGBM works by fitting a gradient-boosting model into the dataset to be trained. While training, a model is made to build decision trees again and again to reduce a specific loss function. It adjusts the parameters of the tree to optimize the model. Finally, in LightGBM, the evaluation is done by counting how the accuracy or mean squared error for classification or rank, respectively, calculates how well the trained model works.

Following applying all required classifiers for prediction, a few techniques were used to improve the classifiers' accuracy. The first technique is a statistical strategy called Principal Component Analysis, in short (PCA), which is used to minimize the number of variables in a dataset. PCA takes a group of possibly correlated variables and breaks them into a smaller set of uncorrelated variables called principal components. The first main component explains as much variation in the data as possible. The second component does the same, taking into account as much of the untreated variability as possible, and so on. Specifically, PCA performs recognition using a linear combination of weighted eigenvectors, or eigenfaces, derived from the covariance matrix of a training set to represent each image of the training set. The recognition is performed by projecting the test image to the subspace of the eigenfaces and the classification is done by finding the minimum Euclidean distance.

Moreover, PCA has the ability to be applied with other classifiers in order to improve their efficiency. The above-mentioned is only beneficial because the classifiers perform better and faster while the dataset is reduced in dimension length due to PCA. Thus, using PCA, many features and noise may be removed from the classifier applied in this research, which will enhance generalization and reduce over-fitting.

The second technique called cross-validation, is a statistical modeling methodology that provides an estimation of the predictive model's intended effectiveness. The dataset is applied to separation into training and testing subsets in order for it to function. The training subset is utilized to develop the model, while the testing subset is utilized to assess the model. This technique is used several times, using various data divisions. The results of each test are combined to assess the efficiency of the model, thus facilitating model selection and evaluation. Cross-validation also helps to evaluate the mean squared error of prediction for the model without the horrors of estimating the error variance. The method, in turn, estimates the

model's MSE, displaces the observations, classifies functions on the training subset, and uses it on the testing subset.

Furthermore, many classifiers, such as those dependent on neural networks, trees, and regression, are using a cross-validation approach. Cross-validation allows one to appraise the model's prediction error externally and ensures neither over-fitting nor under-fitting. The use of cross-validation with different classifiers, as was done in this project, is indeed possible and reliable in disseminating the prediction fault, which aids in selecting a model.

3.2 Universe and Participants

In this project, two datasets were studied and analyzed to predict heart failure. These two datasets included detailed medical records of patients who visited different hospitals for checkups and to seek medical attention. The first dataset, which covered 1177 patients, is considered the largest, while the second dataset, included 918 patients, is less extensive. Each of the datasets provided for this project has many features and information related to each patient. An example for these information, such are the patients' blood sugar, resting blood pressure, creatinine levels, and ages, and many other factors that could be related to heart failure. The two datasets had several patients with differences in their ages.

In the first dataset, there was a vast amount of adult and elder patients, which made the average age reach 74 years old. The focus on older patients could be a result of the importance of age compared to heart failure since it is considered to be one of the risk factors that leads to having heart failure.

The second dataset, on the other hand, had more adults compared to the first dataset, which led to giving an average age of 53 years old. It should be mentioned that the age range of the first dataset did include some young patients since the range was between 19 and 99. On the other hand, the second dataset range was a bit different than the first dataset age range, which only included patients between 28 and 77 years of age. It is crucial to have that vast amount of ages in this study since it is important to know which group of age has more risk of having heart failure, and that would help with predicting and providing beneficial information to prevent future illnesses. It is important to mention that the datasets used for this project were not collected personally. This kind of datasets cannot be collected personally or by surveys since these are critical information. Instead of collecting these datasets by personal surveys

which would make it unreliable and unethical to obtain, the datasets were collected from the hospitals. It is crucial to mention that using surveys to collect this type of data would give unreliable and unprofessional results, which would effect this study. Therefore, using data from hospitals ensures that the information of patients used in this project is reliable and accurate, also it is collected in a secured and safe way.

3.3 Data Collection

3.3.1 Data Collection Instrument

As previously mentioned, both datasets were collected from patients who visited and received medical care in hospitals. As a result, the instruments used for obtaining these datasets were not surveys nor personal means for this project. However, it is essential to refer to the ways used to collect these two datasets. Firstly, the first dataset used in this study was extracted by (Zhou, et al.) from the MIMIC-III database (2021). This database contains records and information for patients who were subjected to treatment in the intensive care unit. As (Zhou, et al., 2021) reported in their article, they used the Structured Query Language queries, in sort (PostgreSQL), and Structured Query Language (SQL) to be able to extract only the patients who suffered from heart failure. They also focused on the patients whose ages were higher than 15 years old, including that these patients were subjected to the intensive care unit. They were also concerned about the missing values, they wanted to obtain a dataset that had fewer missing values, so they made sure that each feature in the data collected had less than 25% missing values to make sure that these missing values would not affect the final results of their study. This would be beneficial for this study as well since fewer missing values will improve the accuracy of prediction. In addition, using PostgreSQL has a powerful impact since it is known for its reliability and stability as an open-source relational database management. Furthermore, not only was the use of PostgreSQL beneficial, but combining it with the use of SQL, which is also a powerful tool, helped the authors to be able to extract a dataset from the extensive database.

The second dataset differs from the first dataset in terms of the data collection process. Different from the first dataset, there were no instruments, not even tools, used for collecting the data. Instead of using any kind of tools to extract the dataset, it was gathered from five different datasets. The credit for collecting and compiling this dataset goes to (Fedesoriano, 2021), who explained that the dataset contains information on patients from five different datasets, which can be found under the Index of Heart

Disease Datasets from the UCI Machine Learning Repository. To create this dataset, (Fedesoriano, 2021) only had to get these five datasets and gather them together, with removing the duplicated information. The idea behind collecting heart disease datasets together is to get one large dataset that can give more accurate prediction when used to predict heart failure like it is used for this project.

3.3.2 Data Collection Procedures

In this project, two datasets were collected and used, and it was essential to know where they came from. Both datasets were provided from a web-based platform called Kaggle. Kaggle is an open-source legit platform that provides datasets and allows data scientists to communicate and collaborate with other users. The platform was used to find the datasets utilized in this project. However, the first dataset, which is owned by (Zhou, et al., 2021) was contained in the massive MIMIC-III database using SQL and PostgreSQL. Zhou et al. were authorized to access the MIMIC-III database, which included 46,520 patients and 58,976 admissions to the Beth Israel Deaconess Medical Center, Boston, USA, intensive care unit. Moreover, by using SQL and PostgreSQL, they had the opportunity to extract the dataset they wanted from that vast database with the criteria they needed for their dataset, that they ended up having a dataset contains 1177 patients records in total. Furthermore, since their dataset was published online as an open-source dataset, a person called (Shahane) posted it on the Kaggle platform (2021). The platform eases access to the dataset for data scientists and researchers who need it for their work and projects.

The second dataset used in this study was collected by (Fedesoriano, 2021). To make this dataset, (Fedesoriano, 2021) gathered five different datasets and made one extensive dataset. These datasets were collected from the Index of Heart Disease Datasets from the UCI Machine Learning Repository. The datasets from the UCI Machine Learning Repository were contained from five different sources, which are the V.A. Medical Center in Long Beach and the Cleveland Clinic Foundation. , University Hospital, Zurich, Switzerland, Hungarian Institute of Cardiology. Budapest, University Hospital, Basel, Switzerland, and Stalog (Heart) Data Set.

The Robert Detrano, M.D., Ph.D. dataset contained 200 observations, while the Cleveland Clinic Foundation dataset contained 303 observations. The Hungarian Institute of Cardiology in Budapest provided a dataset with 294 observations, and the University Hospital in Zurich, Switzerland, together with the University Hospital in

Basel, Switzerland, contained 123 observations. Additionally, a dataset from the Stalog (Heart) Data Set contained 270 observations. After containing these datasets and combining them, (Fedesoriano, 2021) extracted the duplicate observations, which were 272, and that resulted in a total of 918 patient records. The dataset was then posted on Kaggle by (Fedesoriano) to be used in works and projects that are related to heart failure (2021).

As previously mentioned, the two datasets used for this project were collected from the Kaggle platform. Since Kaggle is a widely accessible online source of countless public datasets, which means it is legitimate to use any datasets from this platform. The datasets used in this project were selected according to their reliability and relative to the project. They were carefully selected to ensure that they met the criteria of the project and were suitable for the analysis. The process of collecting these datasets required searching and evaluating all the available datasets in Kaggle to make sure that the chosen datasets were of high quality and would suit the project perfectly. Generally, using the Kaggle platform to get these datasets ensures that this project was built on a trustworthy data source.

3.3.3 Data Analysis Procedures

This section contains the analysis of the two datasets using the Python programming language. Python is the most widely-used and powerful programming language nowadays. It also has one of the most useful libraries and tools which are important for using machine learning methods and algorithms. It was possible to have a deep understanding of the datasets and get more information about them that would be useful for this project.

At first, to be able to start using Python, the necessary libraries were imported into the notebook. The libraries imported in the notebook are the ones which are widely used, for instance, Pandas, Numpy, Seaborn and many others which are necessary libraries that ease data analysis and visualization. More other necessary libraries were imported which are essential, such as Matplotlib, Scikit-Learn, and other libraries which are most used during data analysis. The importance of importing the above libraries has helped perform data analysis and visualization in this project and therefore build models and make the necessary predictions.

After importing the necessary libraries, it was possible to import the two datasets individually in the notebook. By importing the datasets, it was easy to work and explore the datasets, make some statistical analysis, and get an overall idea about the datasets by visualizing them. Moreover, the initial step in analyzing the datasets was to check the shape and size of the datasets and see the type of data these two datasets have. It is essential to determine if the datasets only had integers and float values or whether they contained some string values because if there are some string features, it might be difficult later to see if these features are correlated to the outcome or not.

The first dataset did not contain any string variables, so it needed no changes for its variables. It was possible to check for the first dataset's features and their correlation to the outcome, which is the heart failure events. On the other hand, the second dataset had some string features that needed to be changed to integers or floats. The first feature that had to be changed was Sex, which contained F and M strings, and they refer to the gender of the patients, whether the patient is a Male or Female. It was changed to the binary numbers, either 0 or 1, only to make it easy to see the correlation between Sex and outcome. It should be mentioned that the zero in the new Sex feature refers to the Female patients, and the one refers to the Male patients.

The second feature to change was chest pain, which included four strings, and each string referred to something. The first string is TA, the shorthand for Typical Angina, which was changed to number one in the dataset. It is one type of chest pain that often comes with squeezing, tightness, heaviness on the chest, or even pressure. The Typical Angina kind of pain does not affect a specific spot in the chest; rather, it affects the whole area, and that kind of pain does not last more than five minutes, according to (Gillen, 2022) site. They mentioned that pain will go directly after taking nitroglycerin medicine or stopping the provoking activity. Moreover, the second in the chest pain list, which was changed to number two, is Atypical Angina, in short ATA, which does not have specific symptoms of chest pain; it only comes with nausea, sweating, or weakness, as written in the (Atypical Angina (concept ID: C0741026) - MedGen - NCBI, n.d.). The third type of chest pain, which took number 3, is NAP, which refers to Non-Anginal Pain, which is the kind of pain that is similar to having heart disease, but in fact, it is not. As stated in (professional C. , n.d.), patients with Non-Anginal Pain might feel the pain behind the chest bones. Lastly, the fourth type of

chest pain which is ASY, that comes from Asymptomatic which took number four on the list, that has another name called Silent Myocardial Ischemia, is a pain that comes with the usual symptoms of angina like nausea, diaphoresis, dyspnea, and other symptoms mentioned in (Gul, 2023). They stated that the Asymptomatic condition that brings the feeling of chest discomfort is caused by the shortening of the flow of blood that is rich in oxygen to the heart.

The third feature that was changed in order to assess its correlation to the target was exercise angina, which shows whether patients got angina from exercising or not. The Exercise Angina feature included two values: the first was Y, which referred to yes, the patient got angina from exercising, and the second was N, which referred to no, the patient did not. To make the data more consistent and more accessible to analyze, the values of the Exercise Angina feature were changed to zero as an alternative to no and one as an alternative to yes. It is essential to mention that angina might be caused by exercise, which is the same as it is recorded in the Exercise Angina feature, or it might happen from stress or anything that makes the heart work harder, and it comes like pain in the chest.

The last feature that was changed for this dataset is resting ECG, which had three values: Normal, ST, which referred to the ST to T wave abnormality, and the third value was LVH, which is the shortened word for Left Ventricular Hypertrophy. Resting ECG, or resting 12-lead electrocardiography, is a test that should be done when the doctors suspect that the patient might have abnormalities in the heart, including arrhythmia. This test can be done without surgery, and the results of ECG show the abnormalities and any proof of whether the patient is suffering from left ventricular hypertrophy and coronary heart disease or not. As mentioned before, the results of resting ECG in the dataset included regular, which means that the heart is beating normally at rest. The second value was ST, which included the patients whose results gave abnormality that was either they had a reversal in T wave and/or they had an increase or decrease in ST by >0.5 mV. The last value, LVH, is the record of patients whose results show either a definite or probable left ventricular hypertrophy. The resting ECG values were changed from Normal, ST, and LVH to zero, one, and two respectively, to be able to use it while finding the correlation between the variables in the dataset and the outcome.

After the necessary changes were made to the datasets, they were analyzed and checked whether they included some null or missing values. This check was essential since it would affect the results of the datasets and their predictions. It was found that the first dataset had some missing values in its features, and to be able to work on the dataset, it was necessary to fill in these missing values. There are many ways to fill in the missing values, and one of these methods is to fill the missing values with the mean of the variable. This method is the one that was used for this project since it was checked that each feature did not have more than 25% missing values. Another method was also used to fill in missing values, and it was done on the outcome. Since it had one missing value, it was essential to fill it. It should be explained that the outcome, which is the target, had a binary values and that means it was only zero, which referred to the patients who did not suffer from heart failure, and one, which referred to the patients who did have heart failure. This kind of feature cannot be filled with the used method since that method uses the mean, and the means of the outcome cannot be used. As a result, the mode calculation method was used to fill in the missing value in the outcome. That method calculates the mode, which is the most frequently appeared value in the feature, and fills any missing values with the mode value. In this case, there were more people without heart failure than people who had heart failure in the dataset, which made the mode value zero. On the other hand, the second dataset had neither null nor missing values in its feature, which needed no methods to use at this point.

The second step taken after filling in the missing values was to check the general information of the datasets by using the info method. This method gives a general idea about the datasets; it shows the types of each feature together with the number of features and values for each feature. Moreover, by using the describe method, which allowed users to see more details of the dataset, it was possible to see the standard deviation, the mean, and the maximum and minimum for each feature. After that, the datasets were visualized individually to see each dataset and what it tells. The first step of data visualization was to apply a histogram for each feature to both datasets. The histogram gives a general idea about the feature and shows the type of features, whether they are continuous or categorical. It is important to mention that continuous data means that the data can be any value such as heart rate, weight, and other kinds of features; however, the categorical data is explained by its name, which

means that data falls into a specific group such as sex, diabetes and other features that only have its data grouped.

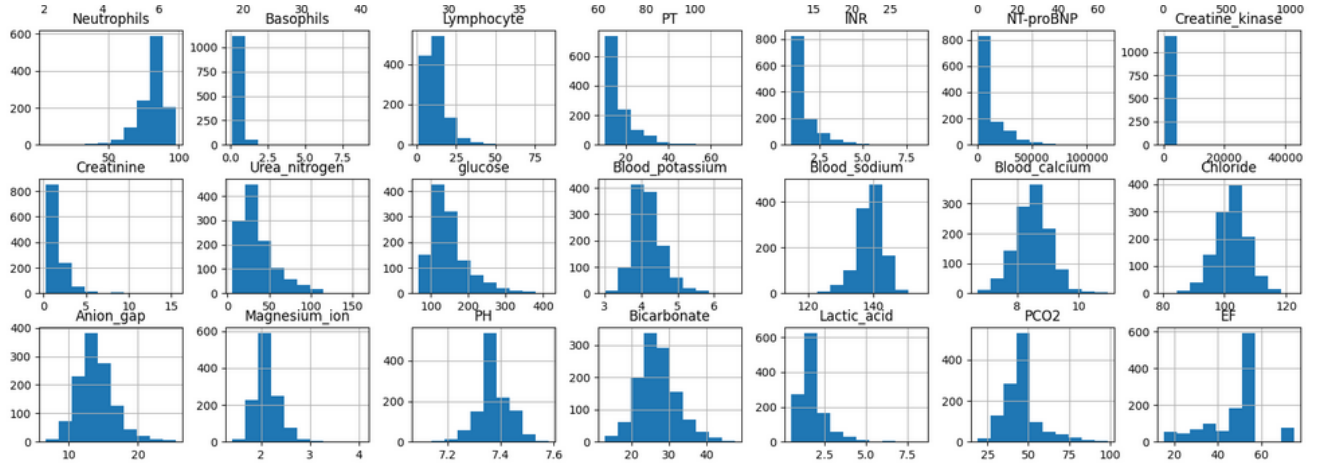


Figure 5 Histogram for dataset 1

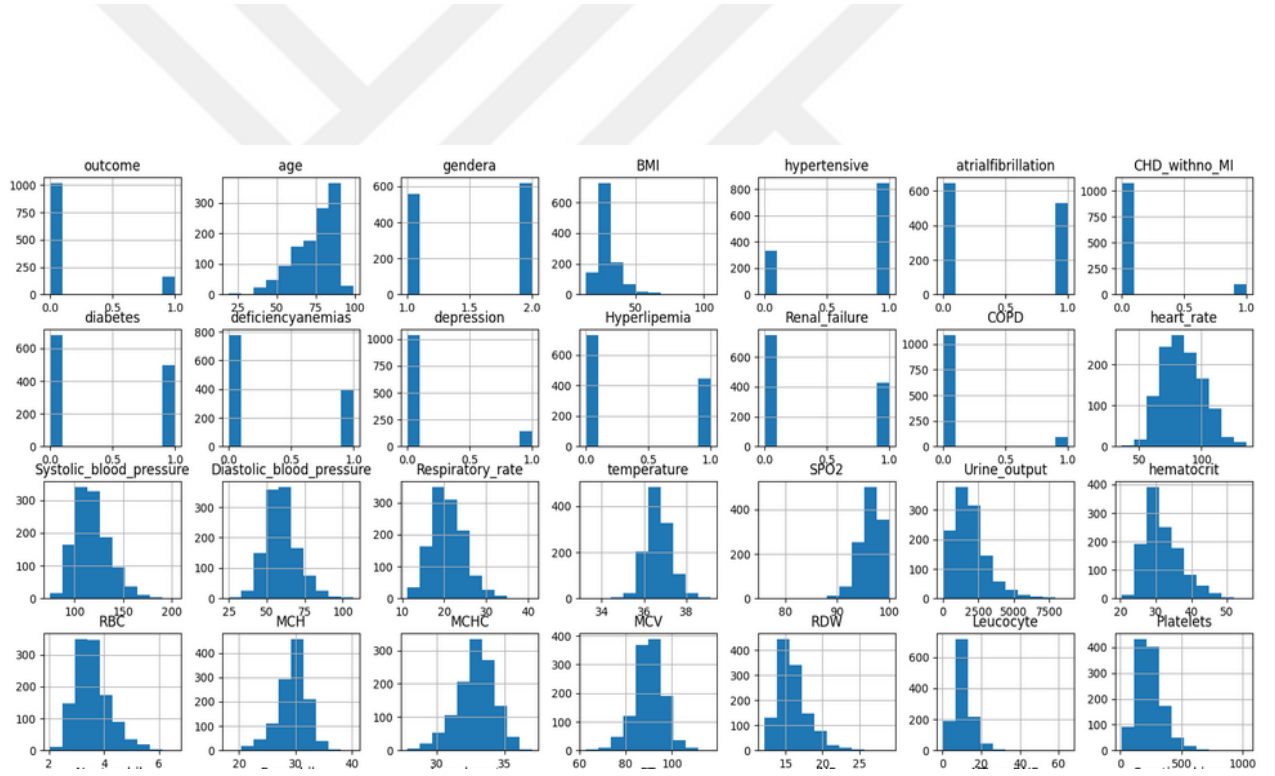


Figure 6 Histogram for the Rest of Dataset 1

The above figures show the results of the histogram for the first dataset, and as it can be seen, it is clear that the dataset had more continuous features than categorical features. Histogram also gave an idea about the categorical features, the outcome as an example, in a way that it was obvious that the patients who did not have heart failure

are more than the patients with heart failure; also for the gender, it is possible to see from the histogram that the number of males is more than the number of females.

The figures below also show the histogram for the second dataset, and they also gave a general idea of the dataset. It can be seen clearly from the histogram that this dataset included both categorical and continuous features equally. Also, it is easy to see from the histogram that this dataset had more patients with heart failure than the first dataset. In general, the histogram plot is vital to get a general idea about the dataset and to see clearly the distribution of the data.

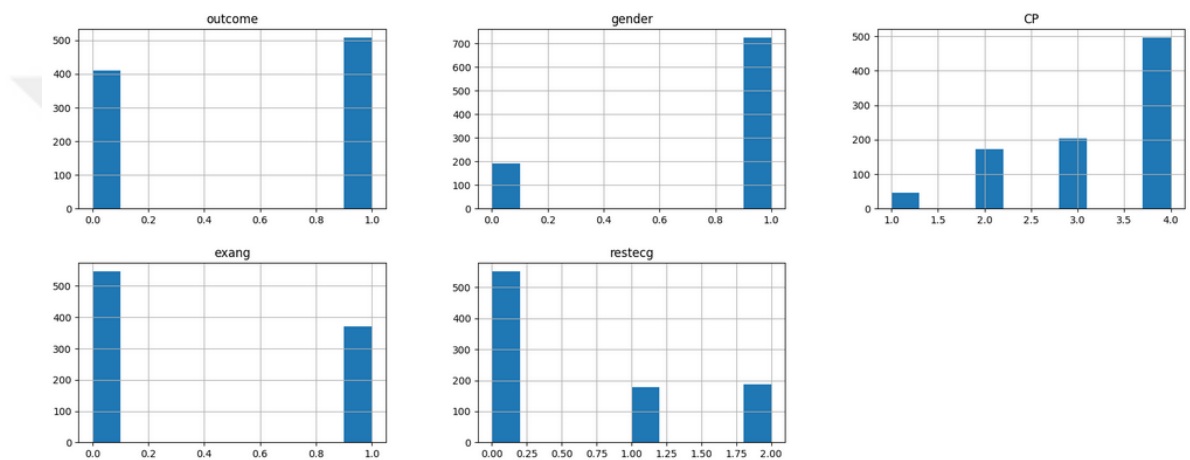


Figure 7 Histogram for Dataset 2

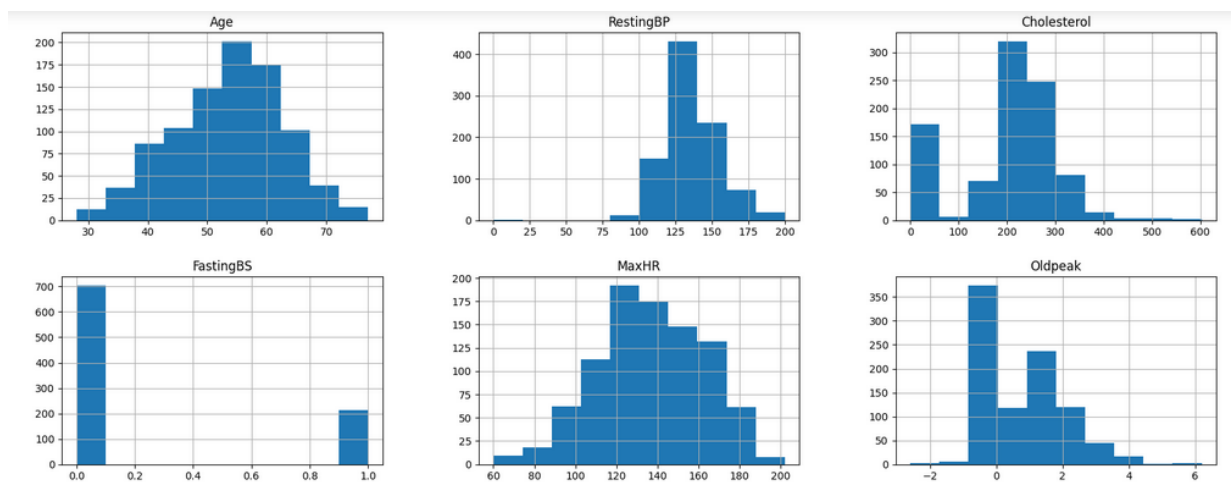


Figure 8 Histogram for the rest of Dataset 2

After the histogram was plotted and gaining a general idea of the datasets, it was essential to check the correlation between each feature in both datasets and the target in the datasets separately. The correlation matrix or heat map is a two-dimensional map that plots all the features in the dataset on the x and y-axis and shows the values of the correlation between all the features in the matrix. The map presents the matrices in different shades of the same color, and each shade shows how much the features they met in that block are correlated. The shades of the heat map tend to change according to the correlation between features that the more the brighter the color, the more it shows the correlation, and the darker it becomes, the less the correlation between the features.

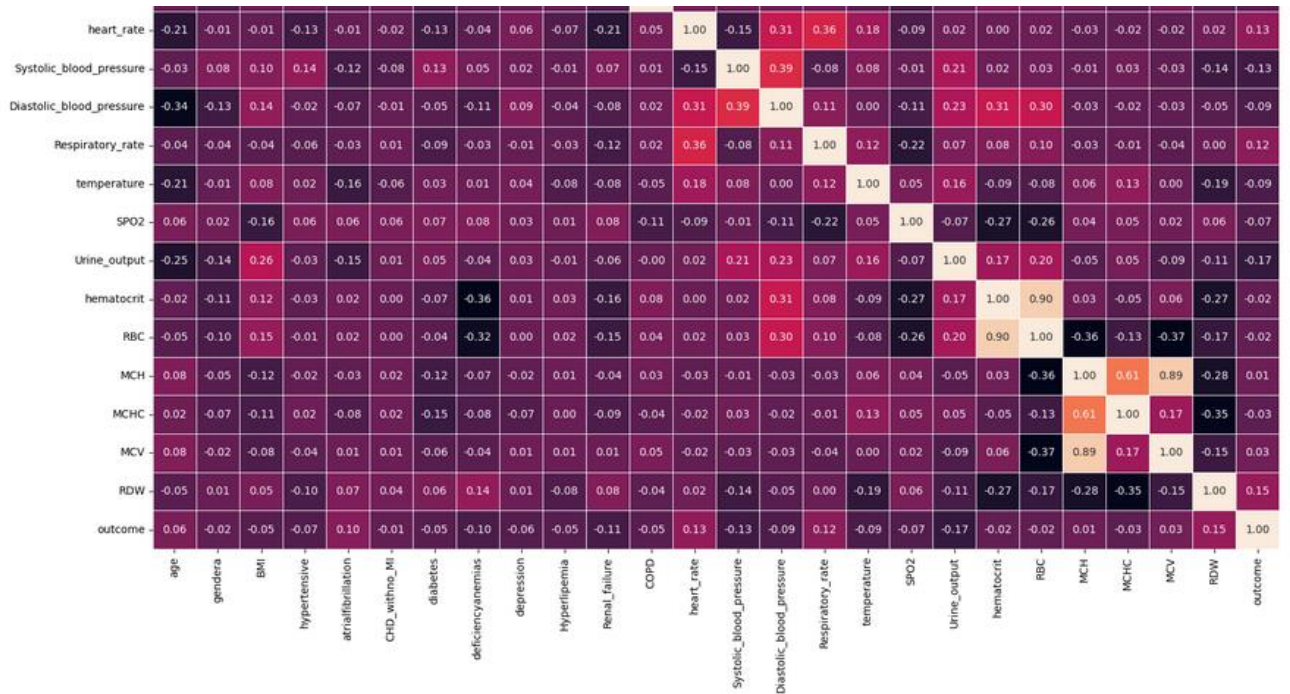


Figure 9 The First Heat Map for Dataset 1

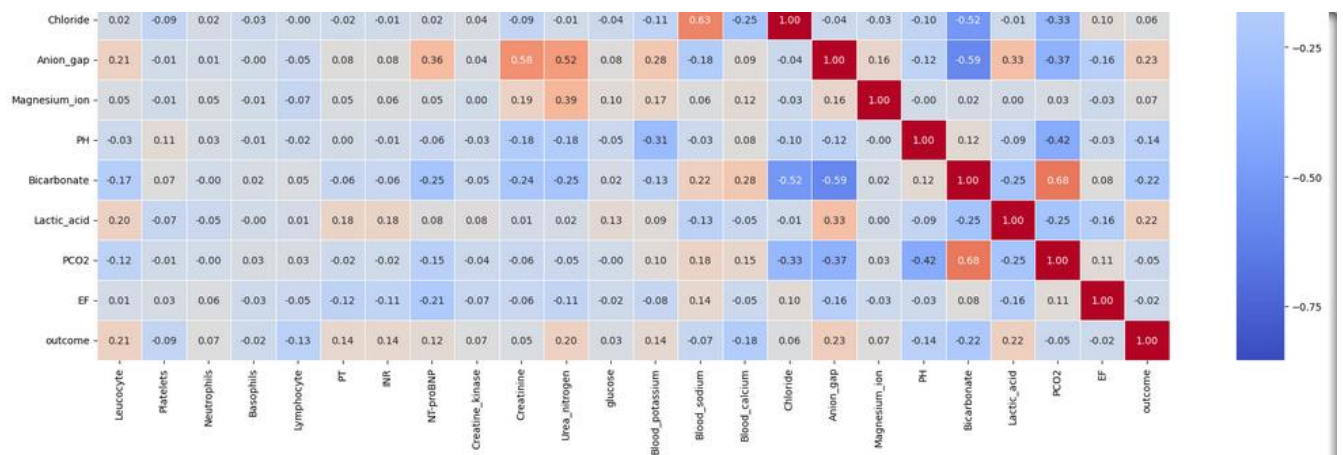


Figure 10 The Second Heat Map for Dataset 1

It was vital to divide the first dataset into two sub-datasets to be able to plot the heat map to it; since the dataset was huge and had 49 features, it took a lot of work to read the heat map when plotting all the features in one heat map. As for dividing the dataset, it was important to include the target in both subsets to get the most correlated features with the target. As shown from the figures above, most features in both subsets showed a correlation with the outcome; for example, from the first figure, it can be seen that the age feature shows a correlation with the outcome, which means that there is a link between the age of the patients and having heart failure. It is noticeable from both figures that the feature that showed the highest correlation with heart failure was the anion gap. It is reasonable that when patients have heart failure, the anion gap increases in their blood.

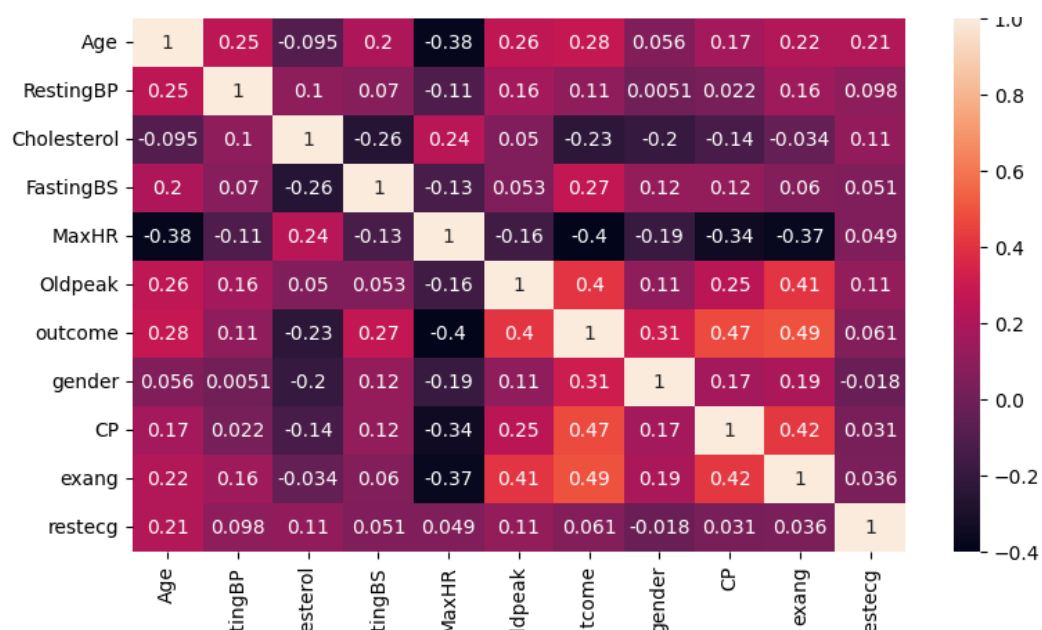


Figure 11 The Heat Map for Dataset 2

The heat map in the above figure shows the correlation between all features in the second dataset and its target. This dataset only has 10 features, which did not need to be divided and made it easier to contain all the features in the same heat map. Moreover, from the clearance of the map, it is possible to see the numbers in each matrix in the map. These matrices represent the number of correlations, and since the critical feature is the target, it is crucial to note that the most correlated feature to the target among the features was exercise angina. It is logical to get a higher correlation from exercise angina since it is one of the critical things that can lead to heart failure.

At the end and after the heat map was used for both datasets, it was important to analyze and visualize the connection between the features and the target. At this point, only the correlated features were applied to data analysis and visualization, and to make it easier to understand, this part was divided into two parts. Since this project uses two datasets, each part of data analysis and visualization presents one of these datasets.

3.3.3.1 Data Analysis and Visualization for the first Dataset

This part is the continuous part of data analysis, including data visualization for the first dataset, which contains 49 features and 1177 of patients' records, and this analysis was made for each feature that showed a correlation with the target in the heat map.

The first essential thing to check was the percentage of patients with and without heart failure since it was noticeable from the histogram that the target of this dataset showed that the number of patients who diagnosed with heart failure was less than the patients who were not diagnosed with heart failure. Therefore, a pie chart was applied to the target, which was helpful to get a clear idea about the target.

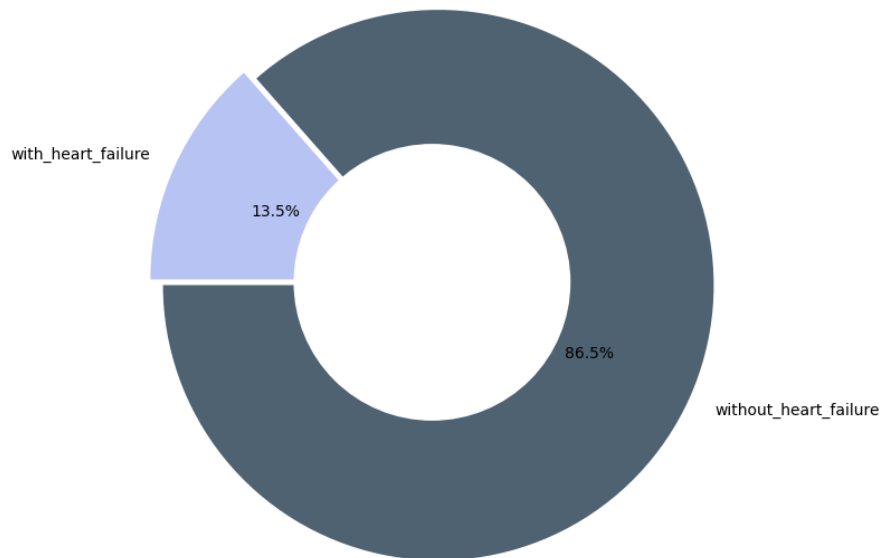


Figure 12 The Pie Chart for the Target of Dataset 1

As shown in the figure above, the pie chart clearly shows the number of patients who have been diagnosed with heart failure and who have not. While analyzing the chart, it is clear that the chart provides the percentage of patients with heart failure, and it was recorded with 13.5% of the total patients recorded.

Furthermore, it was crucial not only to know the number of people diagnosed with heart failure but also the gender of these people, since gender is also considered to be important in this case. Although the gender did not show a positive correlation with the target, which might be a result of the recorded data, it is critical to know the gender of patients because some of the correlated features to the target differ from males to females. Therefore, a similar graphical chart was applied to present the percentage of male patients and females.

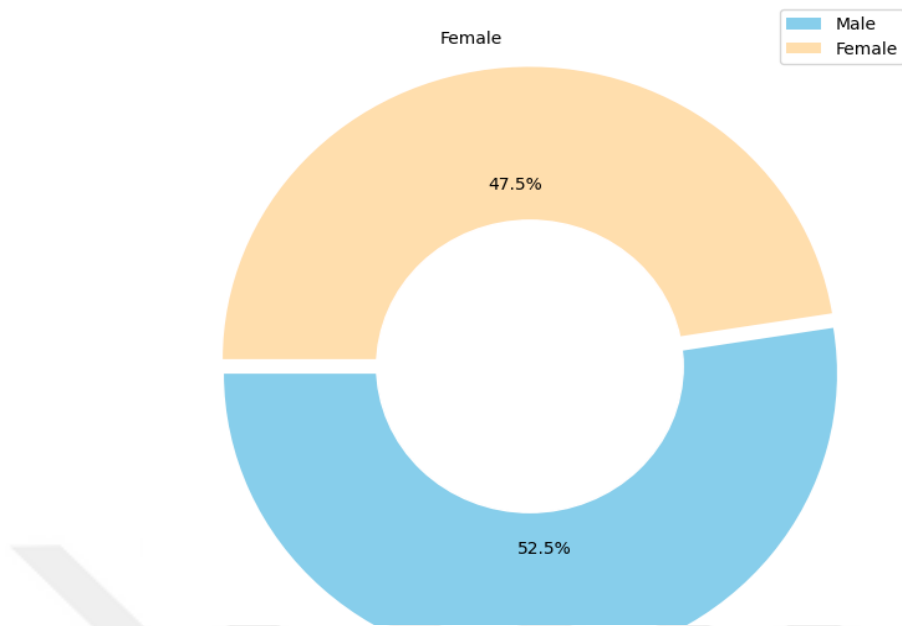


Figure 13 The Pie Chart for the Gender of Dataset 1

The pie chart provided in the figure presents the percentage of males and females in this dataset. It is clear from the chart that the percentage of male patients is more than the percentage of female patients. The chart shows that the percentage of 52.5% was given by males in the dataset, and the remaining percentage, 47.5%, was given from female records in the gender feature.

The feature that was analyzed and visualized after gender is age, and the age feature in the dataset represents the ages of the 1177 patients. The age range in the dataset was between 19 and 99 years old, which made it a bit hard to read it from a pie chart, instead, the bar chart was plotted to find out how the age is correlated to the heart failure, and in which way it effects the patients and their probability of having heart failure.

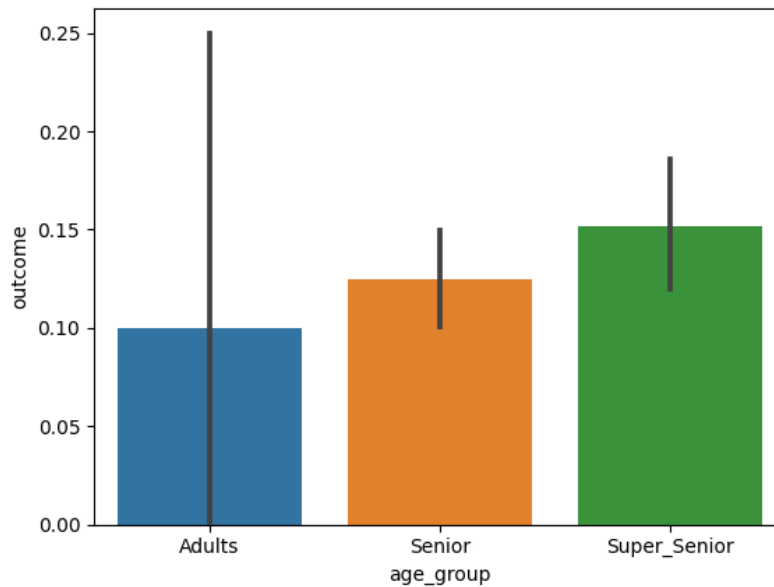


Figure 14 The Bar Chart of Age Vs Outcome

The graphical chart in the above figure is plotted to present the connection between heart failure and age, and as the age range was from 19 to 99 years old, it was important to divide the ages of patients into three groups to give a clear idea of which age range would be in risk of having heart failure. The first group included patients with an age range from 19 to 39, which presents the adult's age in the dataset. The second group range is between 40 and 79 years old, which shows the number of senior patient ages in the dataset. The last group was a super senior group, which represented patients that had ages between 80 and 100 years old. It is clear from the figure above that it was found that the super senior group gave the highest percent among the three groups, and this shows that the number of patients aged between 80 and 100 were more likely to have heart failure. Additionally, by analyzing the age feature, it turns out that the super senior group contains 15.14% of patients from the whole age feature, and according to the figure, these patients were having heart failure. As a result of this analysis, it is possible to say that age has an effect on whether the person is at risk of getting heart failure or not, and it is clear that the older the people become, the more risk of having heart failure they have.

The following feature that was visualized is atrial fibrillation, in short AFib, which is a heart condition that causes some irregular movements in the heart's atria. As stated on the (Atrial Fibrillation, 2024), the irregularity of the heartbeats is caused by the upper atria, which might not show symptoms for some people, but for the others, it can cause shortness of breath, pounding heartbeat, or many other critical symptoms.

This condition could lead to clotting in the heart and could cause heart failure since heart failure is a condition where the heart cannot pump the blood properly.

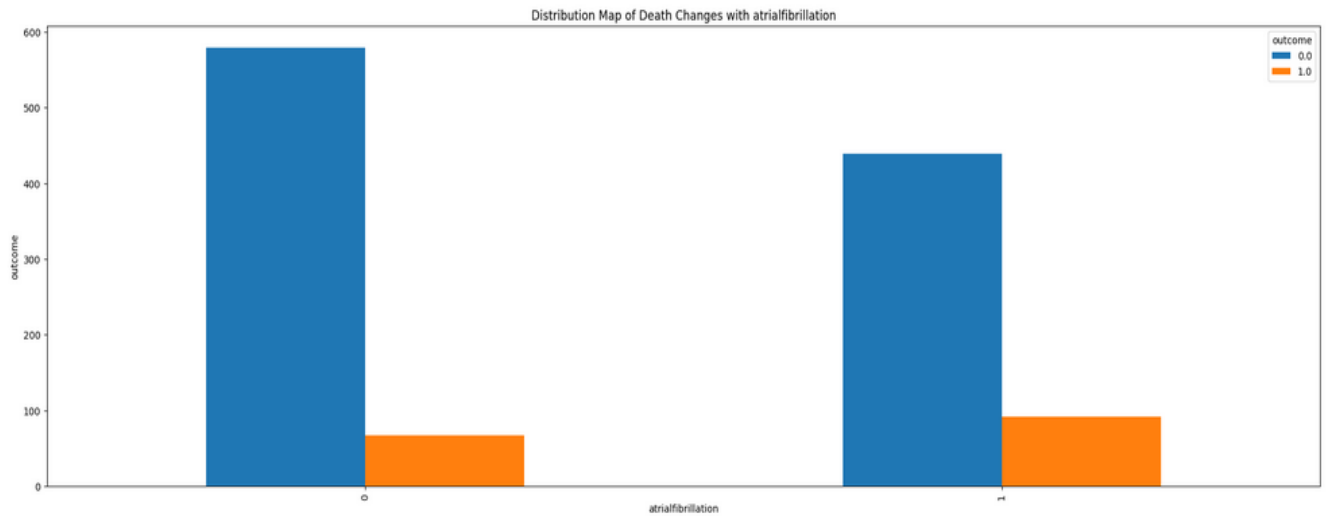


Figure 15 The Bar Chart of Atrial Fibrillation Vs Outcome

The above figure presents the relation between atrial fibrillation and the outcome, which is presented by the bar chart. The atrial fibrillation in the dataset was given as a categorical data type, which means the number zero represents the patients who did not suffer from atrial fibrillation, and one represents the patients who had the condition. Moreover, the chart shows two colors; each color refers to the value of the outcome's values. The blue color refers to the patients who did not have heart failure, and the orange color represents the patients with heart failure. It is shown that people with no atrial fibrillation condition are more likely not to have heart failure, although there was a small group of people who did suffer from heart failure and did not have atrial fibrillation condition. On the other hand, the group of patients who had heart failure and atrial fibrillation condition was more, and this is normal since atria conditions are related to heart diseases in general. It is noticeable as well that there are patients who had an atrial fibrillation condition but did not suffer from heart failure, and this is normal as well, as it was mentioned before that some people might have atrial fibrillation, but they would not show any symptoms.

Following up the analysis by analyzing and visualizing heart rate features, which is one of the most important features related to heart failure. Heart rate is the amount of times the heart is beating at a specific time; normally, it is measured in a minute. While measuring the heart rates, the patient is usually asked to rest to get an

accurate heart rate, as the regular resting heart rate should fluctuate between 60 and 100. Moreover, according to (Edward R. Laskowski, 2022), it is crucial to mention that the normal resting heart rate might differ according to some variables, like being a smoker, fit or active person, and many other variables. However, if the heart rate is constantly higher than 100, the person should seek medical attention because this means that there is a heart condition like heart failure.

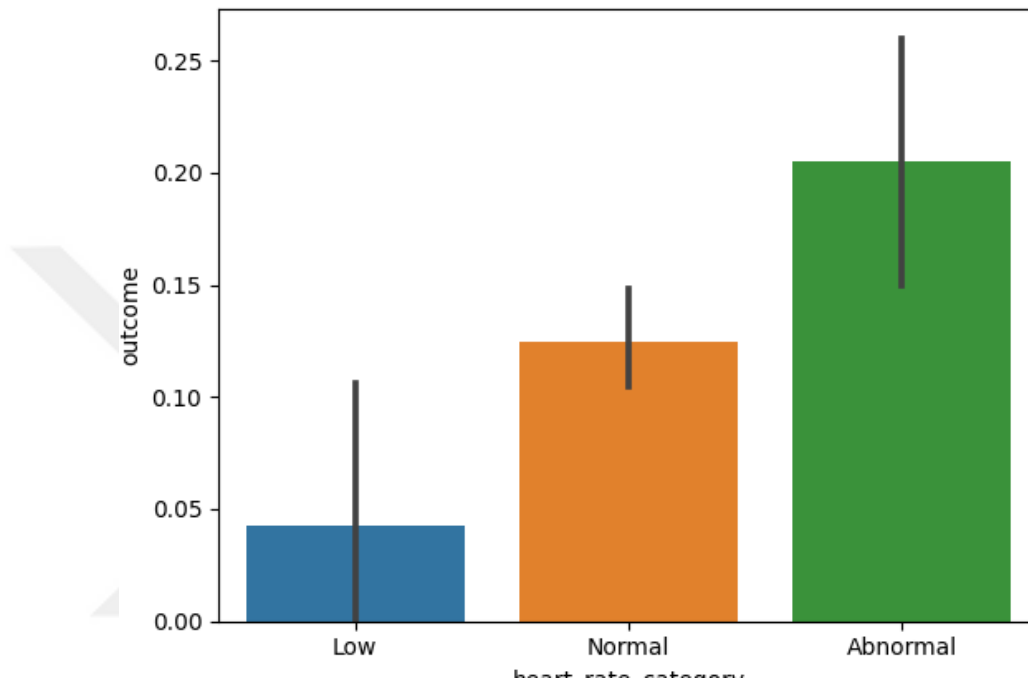


Figure 16 The Bar Chart of Heart Rate Vs Outcome

The figure shows the results of heart rate analysis with respect to the outcome, which is visualized by plotting a bar chart. To make the results clear, it was essential to divide the heart rate records into three groups. While the first group contained patients with low heart rates with a range 30 to 59, the second group had patients who had normal heart rates, which is given the range between 60 to 100, and the last group of patient were having high heart rates that it reaches 136 per minute while resting. The abnormal group, as it is shown in the figure, gave the highest result among the groups, which indicates the number of patients who had abnormal heart rate and heart failure at the same time. This kind of result is reasonable, since the abnormality of heart rate would affect the heart and lead to having heart failure.

The next feature that was analyzed and visualized was the respiratory rate, known as RR, which is the number of breaths a person can take per minute. The

respiratory rate can be a sign of an issue occur in the body such as fever, heart diseases, dehydration and many other problems.

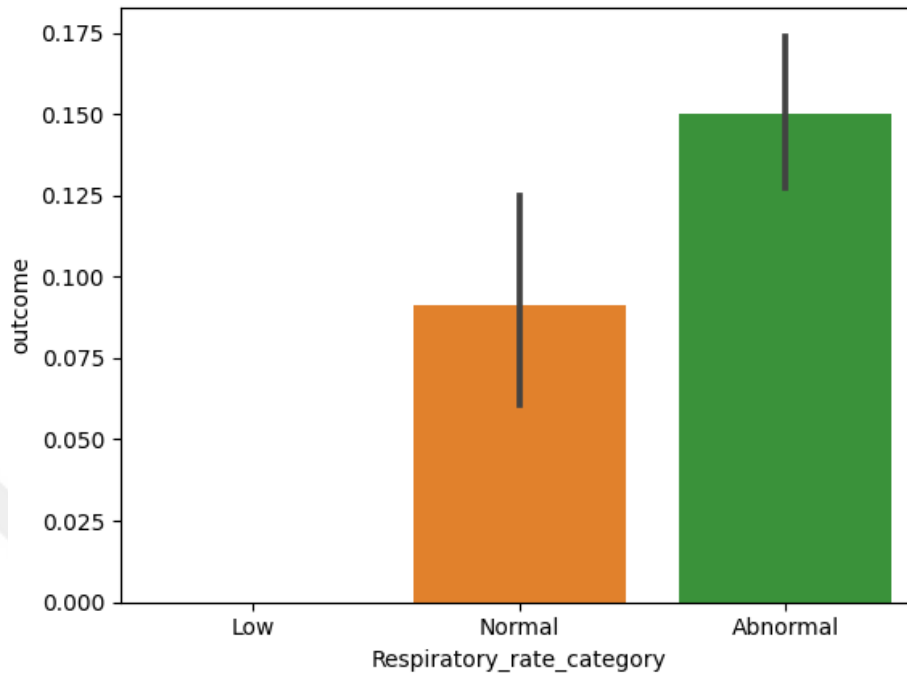


Figure 17 The Bar Chart of Respiratory Rate Vs Outcome

As shown in the chart, the respiratory rate was divided as well into three groups, and each group presented a different level of respiratory rate. The first group might represent the number of patients who had low respiratory rates, and this part of the graph does not contain anything because there was only one patient who had a low respiratory rate. The other two groups, however, showed some results as there were patients who had their respiratory rate normal and others who had their respiratory rate abnormal or high. Usually, a normal respiratory rate has a range between 12 and 18 as number of breaths per minute, and anything above this range is considered to be a high rate of respiratory. Additionally, patients with high respiratory rates showed higher levels of heart failure than the other groups, which is sensible since the increase in respiratory rate is a sign of an issue with the person, and that would be caused by heart failure.

The following analysis was on the feature MCV, which refers to mean corpuscular volume. The mean corpuscular volume present the volume of the red blood cells in the body, the change in its volume would affect the body in many ways and it is also considered to be a sign of issues occurred in the body. It was essential to

visualize the MCV since it has a direct relation with the heart failure and it also showed correlation with the outcome.

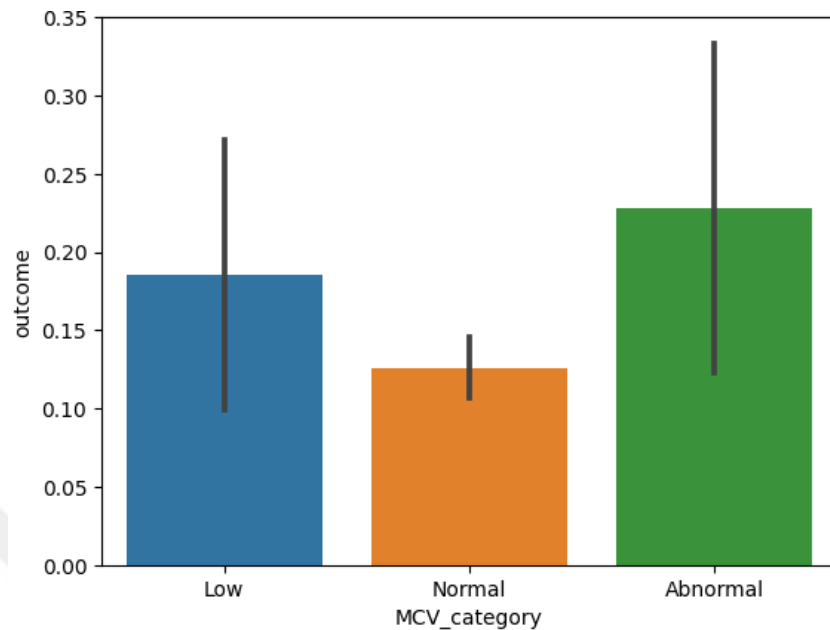


Figure 18 The Bar Chart of MCV Vs Outcome

The chart is plotted to visualize the relation between the mean corpuscular volume and the outcome, and as the records of the patients included high and low MCVs, it was important to divide the feature into three groups. The usual volume of red blood cells is between 80 and 100, and if the blood test come with a result of MCV lower than 80, it might indicate a kind of anemia. As well as, if the results come with high MCV, as mentioned in the (professional C. , Mean Corpuscular Volume (MCV) Blood Test V, n.d.), it is considered as a sign of a medical problems such as liver diseases, vitamin B12 deficiency, and other medical conditions. Moreover, from the figure, it is clear that patients with low and abnormal or high MCV were at risk of having heart failure more than patients with normal MCV. The figure also shows that patients with high MCV had higher results of having heart failure than patients with low MCV, which is convenient since the heart failure could cause an expansion in the blood and that indicates an increase in the mean corpuscular volume.

The next feature visualized was RDW, which is a shortened name for red cell distribution width, a blood test can measure that. As mentioned in (RDW (Red Cell Distribution Width): Medlineplus Medical Test, 2022), the red cell distribution width in the blood test calculates the changes in size and volume of the red cells, which is important

to check since the imbalance in the RDW would lead to some health problems. Additionally, the type of health problem changes according to the RDW levels, as the lower level of RDW would indicate that the patient might have iron deficiency, different types of anemia, and other conditions. On the other hand, higher levels of RDW could be a sign of different conditions like diabetes, kidney diseases, heart diseases, and other critical conditions. It is crucial to mention that the normal range of RDW differs for males and females, and because of that, the RDW feature was divided according to gender, and each group was visualized individually.

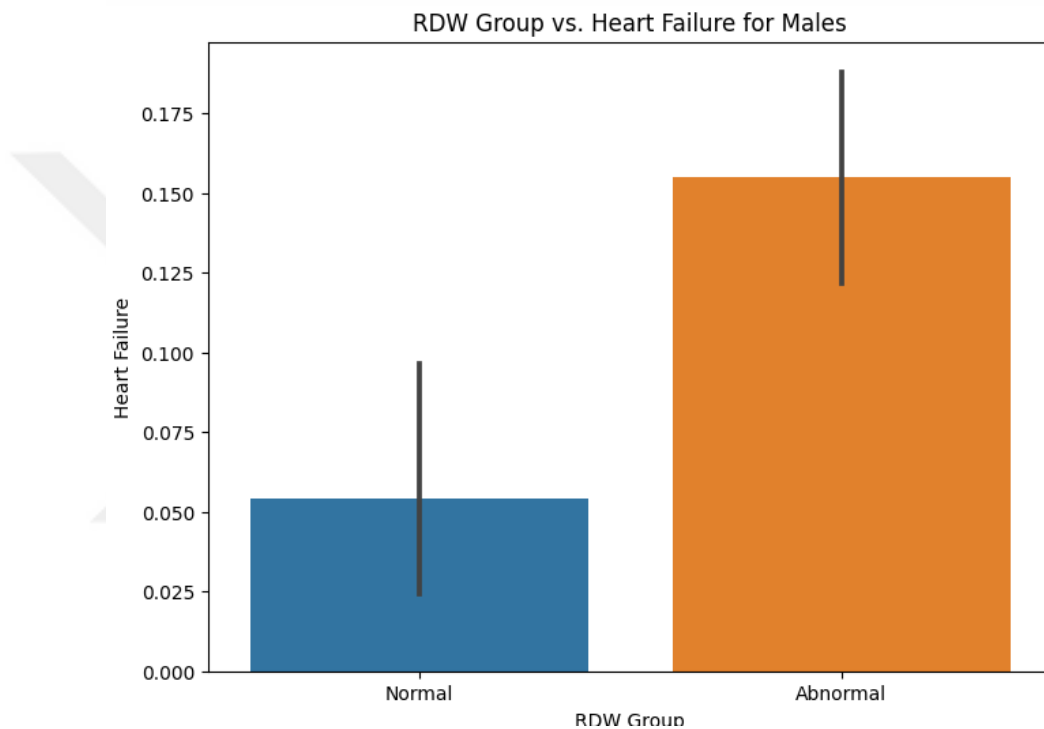


Figure 19 The Bar Chart of RDW Vs Outcome for Males

At first, the RDW of male patients was visualized as shown in the figure, and it was important to check the normal and abnormal ranges of RDW for males. The normal range for males is between 11.8 and 14.5 percent of RDW, which is presented by the blue bar in the figure, and any records above or below this range are considered to be abnormal and are presented by the orange bar. Furthermore, it is clear from the figure that male patients with an abnormal level of RDW showed a higher risk of having heart failure than patients with an average level. It is essential to mention the analysis done to this part showed that 15.97% of male patients with abnormal RDW were recorded as heart failure patients.

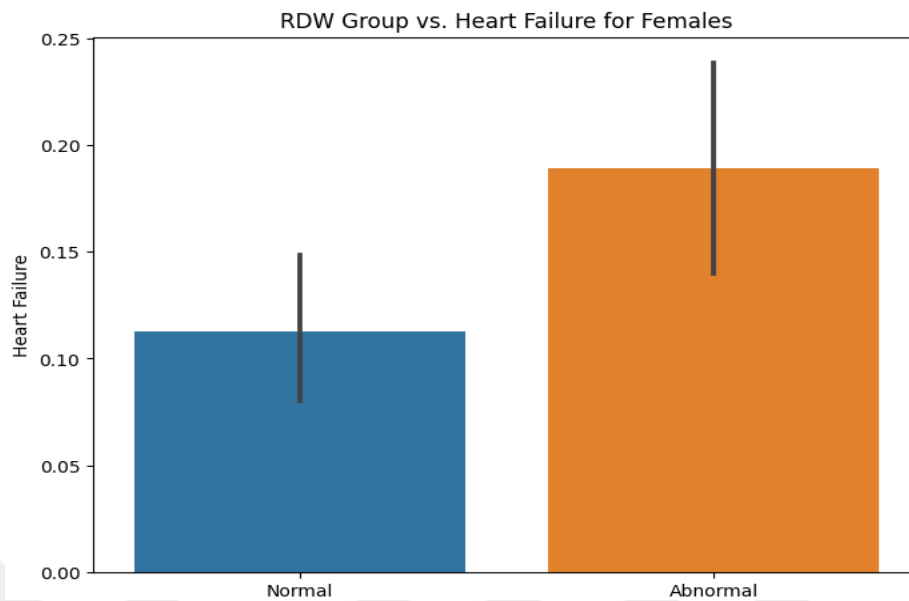


Figure 20 The Bar Chart of RDW Vs Outcome for Females

The above figure shows the visualization of RDW for the female patients, and it was also considered the normal and abnormal ranges for females. The normal ranges of RDW for females are between 12.2 and 16.1 percent, which is the blue bar in the figure, and other female RDW records that were more than 16.1 or less than 12.2 are considered as an abnormal percent. In addition, after analyzing the number of females who had an abnormal RDW, it turned out that there is 19.69% of the females who had an abnormal RDW and suffered from heart failure at the same time. Although the records of female patients are less than the male patients in this dataset, it is quite vital to consider the results of female patients who had both heart failure and abnormal RDW compared with the male patients who had the same. The results of the females were higher than the males, which would give an idea that females might have a higher risk of having abnormal RDW and heart failure. Moreover, it is important to mention that the effect of having high RDW is the same for females, and containing a healthy and normal RDW is as important as it is for males.

Another feature was analyzed and visualized, and it is Leukocyte, which is also known as white blood cells. It is a part of the blood that fights diseases and protects the body from all unusual bodies entering the human body. As a result, it is important to check Leukocytes since the changes in them would show what kind of health problems the patient has. As stated on (professional C. , Low White Blood Cell Count (Leukopenia): Causes, Symptoms & Treatment, 2024), the decrease in Leukocyte

could indicate the patients have a kind of autoimmune disorder such as lupus, or they might have gotten some treatments that caused this drop in Leukocyte such as cancer drugs. However, having an increase in Leukocyte could indicate other issues and infections. Some of these problems could be everyday things such as having an injury or getting burned, which would increase Leukocyte. However, other issues might indicate serious problems such as leukemia or heart disease like heart failure, as mentioned (Madjid & Fatemi, 2013).

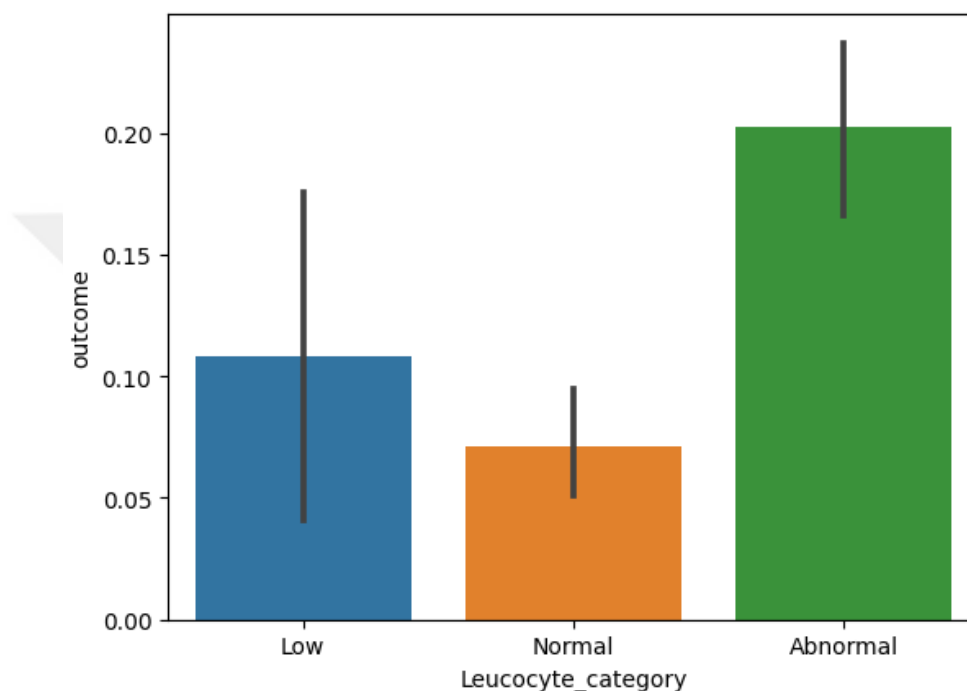


Figure 21 The Bar Chart of Leukocyte Vs Outcome

Based on the bar chart shown in the figure above, it was evident that patients with regular Leukocyte levels did not have the risk to develop heart failure with a comparison to the high and low Leukocyte levels. It is worth noting that the patients with low Leukocyte levels showed some relation with having heart failure, even though usually having low levels of Leukocyte is not directly related to developing heart failure. On the other hand, it can be seen that most patients with high levels of Leukocytes seemed to have a relation of having heart failure, and this is considered a reasonable result since the increase in Leukocyte levels is directly affected by the risk of having heart failure. In general, the analysis done to study the relation between Leukocyte levels and heart failure showed reliable results.

Another critical feature was analyzed, which is PT, which is the shortened name of prothrombin time. The prothrombin time is a test that calculates the average period of time that the human body takes to heal a wound, in other words, to clot the blood. The usual time the body takes to create blood clotting is between 11 and 13.5 seconds. Any number change in the time of clotting would induce an issue in the body, as the increase in clotting time indicates that the patient might suffer from a serious injury in the liver, and it might be a sign of a damaged liver. Moreover, it is stated in (Viral Hepatitis and Liver Disease, 2018) that the liver is the organ that is responsible for creating the proteins, which make clots when bleeding. As well, high prothrombin time indicates that the liver is not producing enough of these proteins, and that leads to taking more time for the blood to clot, which means that the liver is not in a good condition. As a result, this damage would affect the heart as it might be a cause of developing heart failure.

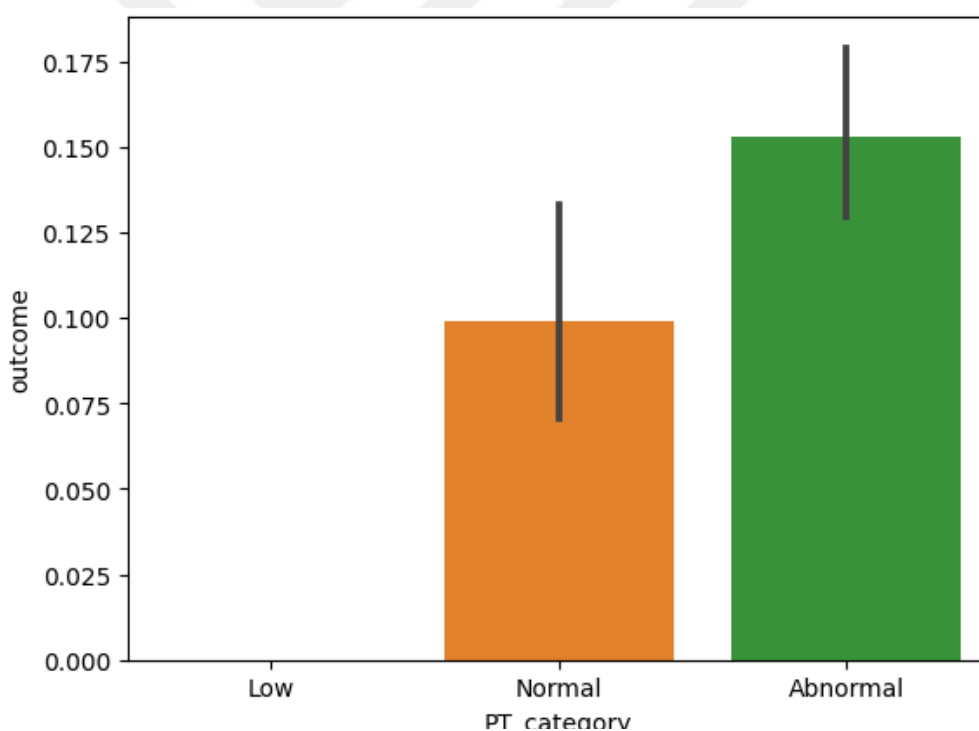


Figure 22 The Bar Chart of PT Vs Outcome

Regarding the above figure, which includes the bar chart of PT with respect to heart failure, it is clear that the records of patients were divided into three groups. The first group, which represented the patients who had low prothrombin time, did not record any changes, and this was a result of the number of patients, which included only two patients with low PT. It is essential to mention that it would not be reasonable to add these two to the normal group since these patients did not have an average

prothrombin time, and including them in the normal group would affect the overall analysis results. Furthermore, the second group, which includes patients who recorded standard prothrombin time, showed some relation with heart failure, and that might be a result of suffering from other health issues that would be related to heart failure. The last group, which included people with high prothrombin time, which had the range of 13.4 to 71.2 seconds of clotting time, provided a higher correlation with heart failure than the other two groups. This correlation is logical since the increase in prothrombin time is a cause of liver damage and that would affect the heart.

Following the analysis of PT, the next step was to analyze INR, which is the shortened name of the international normalized ratio. The INR is basically a kind of calculation based on prothrombin time, and it indicates the same reasons as PT. The main reason for determining the INR level is to find if the blood is having a normal time of clotting or not. It is vital to know that the higher level of INR indicates that the person might be at an increased risk of bleeding. This is because an increase in INR level indicates that the blood is taking more time to clot, which would be a severe condition, especially for people who already suffer from health problems that can be affected by the blood clotting process. As it is in PT, it is crucial to maintain a normal level of both PT and INR because the abnormality for these two is critical and would affect the heart.

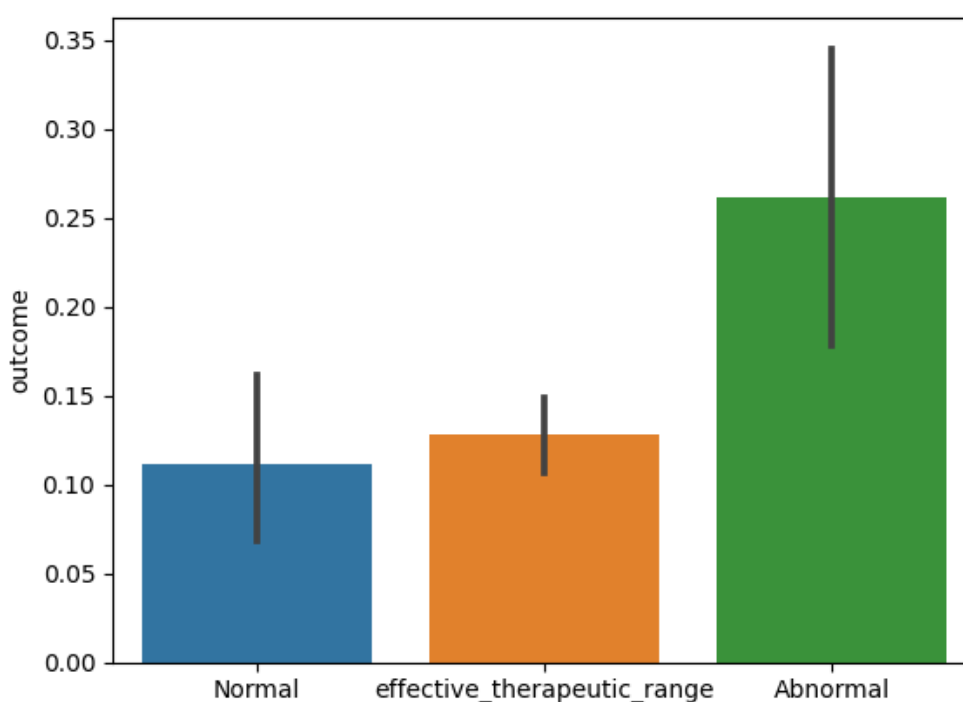


Figure 23 The Bar Chart of INR Vs Outcome

According to the figure above, which presents the relation between INR and outcome, it is clear that the patients in the dataset had several levels of INR. As a result, it was crucial to divide these levels according to the normality and abnormality of the INR levels. The first group contains the patients with normal INR levels, which are maintained between 0 and 1.1. The second group had a number of patients with 2.0 to 3.0 INR levels; this was specified because people with this level of INR consider it as an effective therapy for them since they are taking warfarin for some health conditions, as stated in the (Prothrombin Time Test, 2022). These conditions might include blood clotting in the lungs or in the legs. Also, it might be taken for atrial fibrillation. The last group included patients with higher INR, and this group showed a higher relation with heart failure since the higher levels of INR increase the risk of bleeding, which, as a result, increases the risk of having heart failure. The results of this analysis were reasonable since it gave almost the same results as the one recorded from PT.

The next analysis was for Creatine Phosphokinase, which is also known as Creatine Kinase. It is a protein, in other words, an enzyme that is located in the person's brain, muscle tissue, and heart. This enzyme is found when a damage occurs in the heart or muscle tissues, so when it is found in the blood with high levels, it indicates that an injury happened to the person's heart or muscles that made Creatine Kinase leak into the blood. Additionally, it is also known that finding CK in the blood might indicate some other problems like having heart failure, as mentioned by (Creatine Phosphokinase (CPK), 2019). The site explains that it is highly likely to find high levels of CK after doing some workouts or even after taking some specific medications. As a result, it is essential to measure CK during regular blood tests to ensure that the body does not have an inside injury or have a risk of getting heart failure.

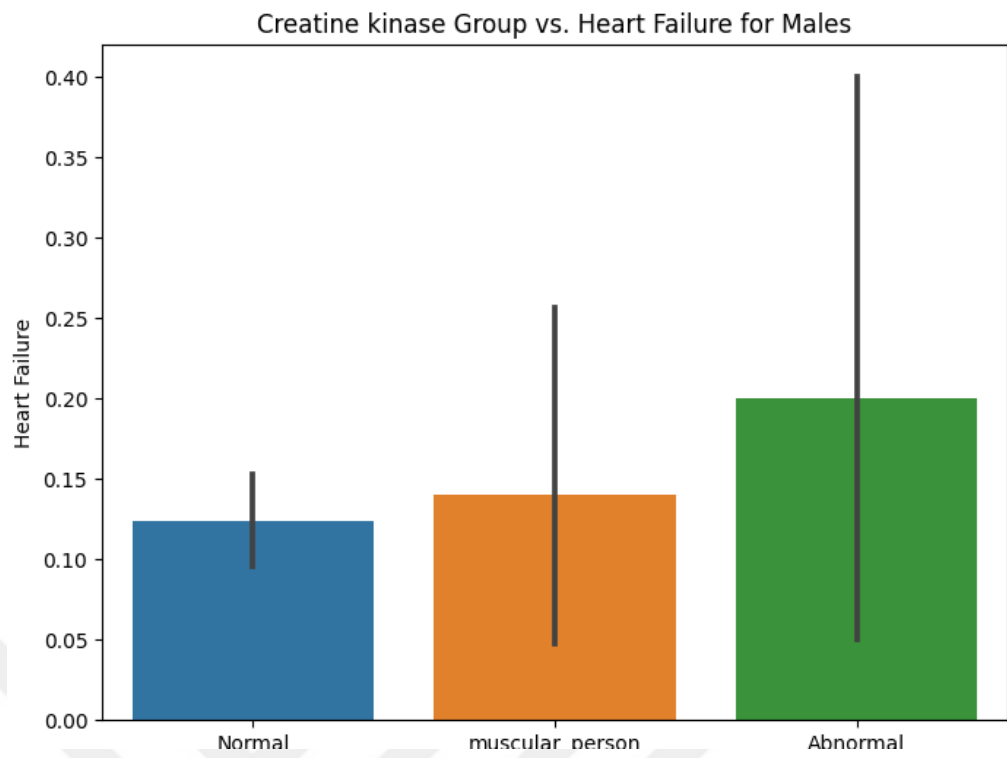


Figure 24 The Bar Chart of CK Vs Outcome for Males

It is important to mention that the normal levels of CK in blood for males differ from the normal levels for females, and as a result of that, the records of creatine kinase were divided according to the patient's gender. The figure above presents the relation between creatine kinase for males and outcome, and it is clear that the records of males included different levels, which made it vital to group the levels according to normality and abnormality. The first group included patients with normal CK levels, which is from 40 to 320 units/L, and the second group is considered to have normal levels for muscular people, which is from 321 to 712 units/L. The third group, however, contained patients with high levels of CK, and it is the one that showed a higher correlation with heart failure. In addition, the analysis made for the male patients who had abnormal levels of CK gave out the percentage of patients with heart failure, which was 23.8% of male patients having high levels of CK with heart failure. These results are reasonable since there is a clear relation between problems in the heart and the increase in the level of creatine kinase.

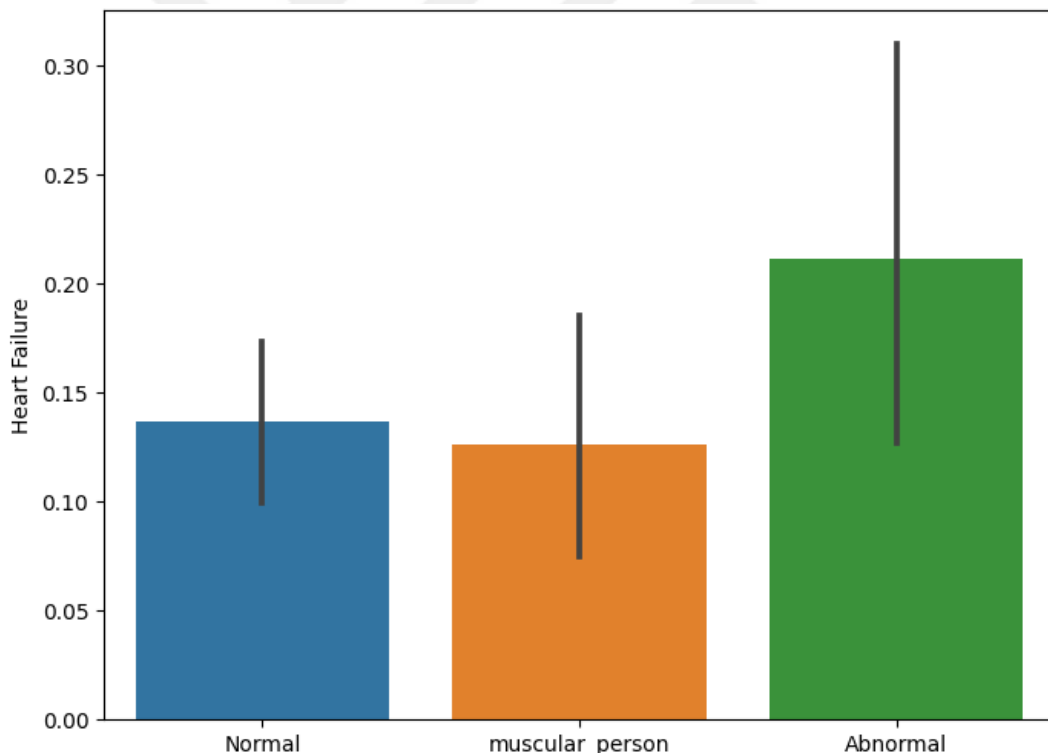


Figure 25 The Bar Chart of CK Vs Outcome for Females

The bar graph in the above figure shows the visualization of the outcome and the creatine kinase level of females. The female patients in the dataset provided similar changes in creatine kinase levels, which also had to be divided into three groups. It is crucial to mention that there are differences in normal levels of creatine kinase in males

and females, and that made the first group of females, which is the normal group, have a range of levels between 20 and 200 units/L. The second group for females also indicates the muscular person, but with a different range of CK levels from males, which was recorded between 201 to 330 units/L, and similar to the male muscular person group, the female muscular person group is considered normal as well. Lastly, the third group shows the females with abnormal CK levels, and as it can be seen that this group gave the highest results among the groups, and that indicate that the abnormal levels of creatine kinase for both males and females could be a sign of higher risk of heart failure.

The feature visualized after CK was creatinine, and it is a product that should not be maintained in the blood, and is coming from muscles after failure or comes from digesting the protein that comes from the food. This product is supposed to be drained out of the body through the kidneys. However, finding creatinine levels in the blood could be normal unless it accedes to the maximum level. According to the maximum level of creatinine in the blood could be problematic, as it indicates that there would be some issues with the kidneys. The issue caused by high levels of creatinine could result in issues with the heart as well. In addition, the heart might be affected when the kidneys are not functioning well, and that is due to their ability to balance fluid in the blood and regulate blood pressure. As a result, it is essential to measure the creatinine levels in the blood regularly to ensure that the kidneys are working as well as the heart.

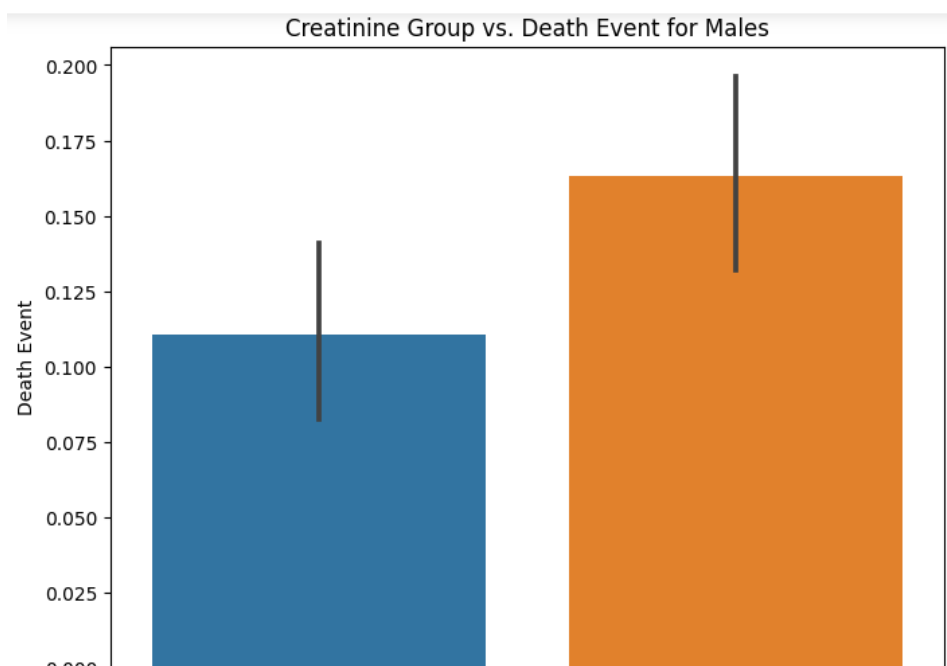


Figure 26 The Bar Chart of Creatinine Vs Outcome for Males

Based on the figure above, which shows the relationship between creatinine levels for male patients and outcome, it was crucial to separate the male patients and females to get a clear reading for creatinine levels. The separation happened because of the differences in minimum and maximum levels of creatinine for males and females. From the above chart, it is clear that male patients had normal and abnormal levels of creatinine, which made it important to have two groups to get a detailed analysis. Moreover, the first group in the graph contained male patients with normal levels of creatinine, which is between 0.74 and 1.35 mg/dL. The second group, which contained the male patients with an abnormal creatinine level, is the one that gave a higher correlation with heart failure. This result is considered normal and logical since high levels of creatinine could cause heart failure.

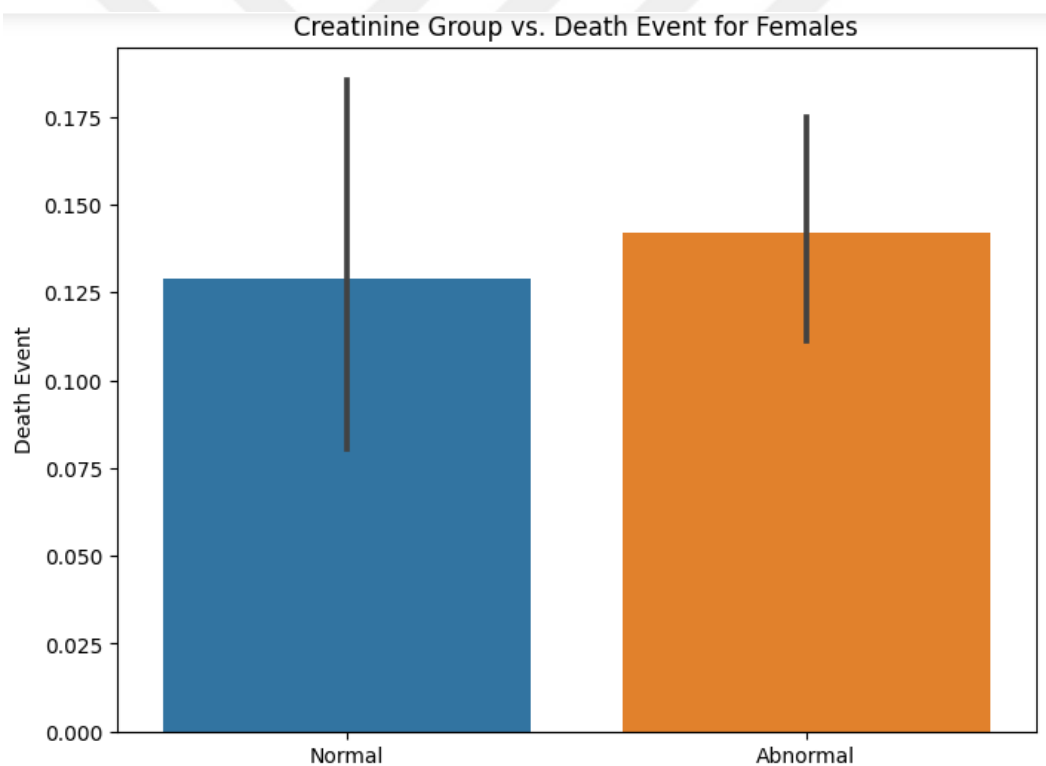


Figure 27 The Bar Chart of Creatinine Vs Outcome for Females

The figure above presents the relation between outcome and the creatinine levels of female patients, and as it was done to the males group, the female patients group was also divided into two groups. The first group for creatinine levels of females is the one that includes female patients with normal creatinine levels, and it is

considered to be between 0.59 and 1.04 mg/dL. The second group presents female patients with abnormal creatinine levels, which is why it has a high correlation with heart failure. In addition, it should be mentioned that compared to the male patients who had abnormal creatinine levels and heart failure, there were not many female patients who had both heart failure and creatinine abnormal levels. This might be a sign that males might have more risk of having high levels of creatinine and heart failure than females. On the other hand, females who had normal levels of creatinine were at risk of having heart failure more than the males with normal levels of creatinine, and this result might be caused by an effect of other factors.

Another feature that was analyzed was urea nitrogen, which is also a product that the body flushes out through the kidneys. Urea nitrogen is also developed when eating proteins, and it goes to the kidneys by the blood. After that, the kidneys filter the blood and get rid of these products. This product can be determined by a blood test, which shows the level of urea nitrogen. Moreover, if the level of urea nitrogen is higher than the normal levels, it could indicate an issue with the kidney, which might lead to heart failure. As stated by (Duan, Li, & Yang, 2023), the increase in the urea nitrogen levels, even by 10 mg/dL, could lead to a massive increase in mortality for people with heart failure; it might even reach 21% in increments.

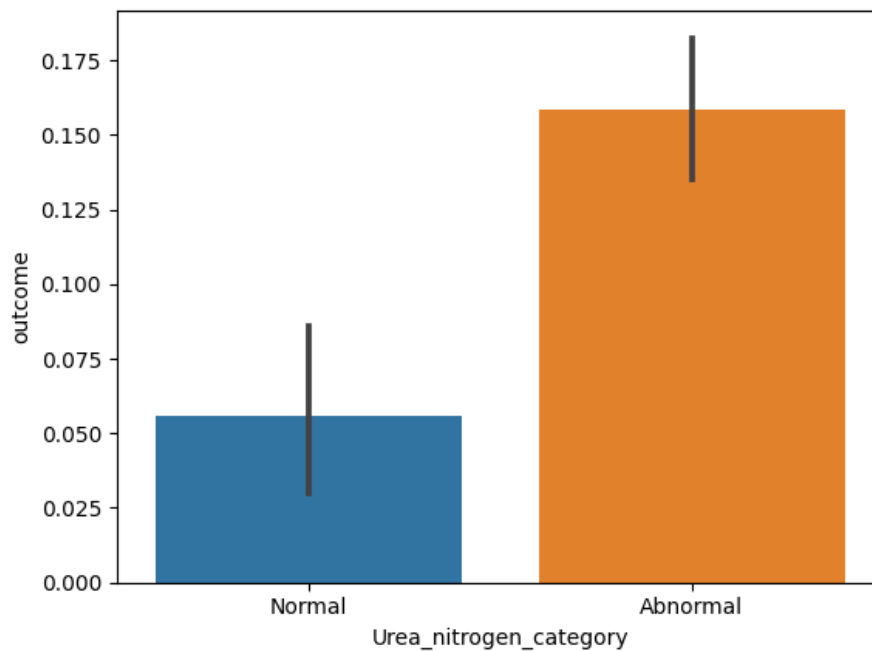


Figure 28 The Bar Chart of Urea Nitrogen Vs Outcome

Regarding the above figure, which shows the analysis done on the urea nitrogen with respect to the outcome, it was evident that some of the patients in the

dataset had high levels of urea nitrogen in their blood. As a result, the analysis included two groups, and each group included a specific part of the patients. The first group, which include the part of patients that had normal levels of urea nitrogen in their blood, and that is a range which starts at a minimum level of 5 mg/dL and ends at a maximum level number which is 20 mg/dL. The second group, which includes the part of patients that had an abnormal level of urea nitrogen in their blood, and it is the group that showed a higher relation with heart failure. During the analysis, it was found that the percentage of patients who had an abnormal level of urea and heart failure at the same time reached 15.84 percent. Furthermore, the results of the analysis made a clear idea of how urea nitrogen levels in the blood would affect the heart, and from the results, it could be seen that most of the patient with high levels of urea nitrogen in their blood ended up suffering from heart failure.

The feature analyzed after urea nitrogen was the blood potassium, which indicates the results of a blood test for knowing the potassium levels in the blood. Potassium is considered to be an electrolyte, which helps of controlling the pH balance. It helps in controlling nerve and muscle activities as well as helping in balancing the fluid levels, as explained by the (professional C. , Low Potassium Level Causes (Hypokalemia), n.d.). Additionally, the potassium level in the blood has a specific range that should not be more or less than that range, this range is considered between 3.5 and 5.2 mmol/L. People who get blood potassium results between 3 and 3.5 mmol/L are considered to have mid-hypokalemia, and people with hypokalemia could face some symptoms such as heart palpitations, tingling, extreme fatigue, and muscle weakness. On the other hand, people who have blood potassium levels of more than 5.2 mmol/L are considered to be hyperkalemia. According to (Simon, 2023), hyperkalemia could also lead to some severe conditions like paralysis or heart arrhythmia.

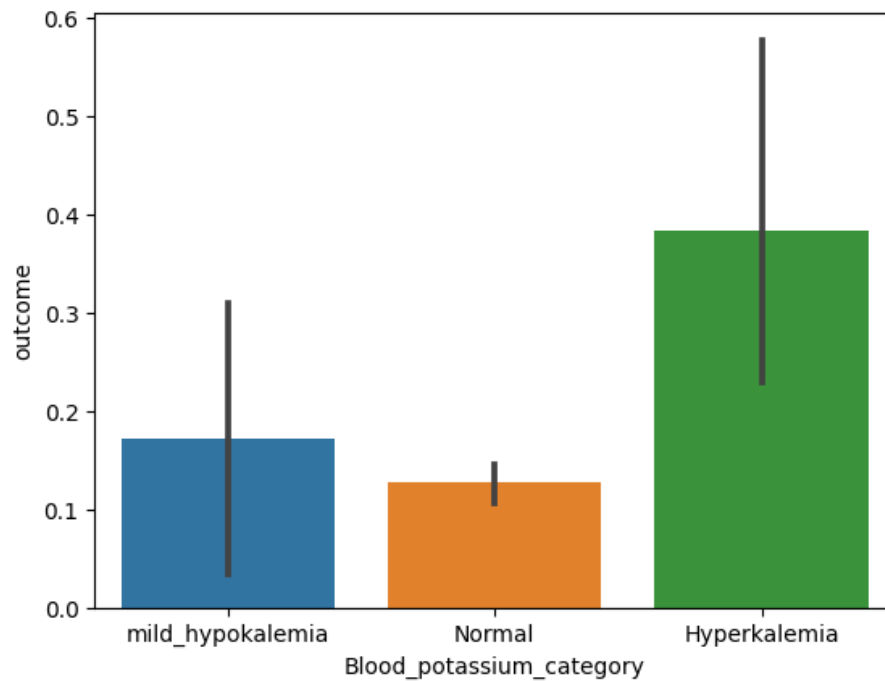


Figure 29 The Bar Chart of Blood Potassium Vs Outcome

Based on the bar graph in the figure above, the analysis was made on the blood potassium with respect to the outcome, and as shown in the figure, the feature was divided into three groups, and each group contained a specific number of patients according to their blood potassium. The first group presents the number of patients who had their blood potassium levels between 3 and 3.5 mmol/L, and this amount of potassium in the blood is called mid-hypokalemia. The second group presents people with normal levels of blood potassium, and the last group depicts the patients who have high levels of potassium in their blood. As a result of the analysis, it is clear that patients with hyperkalemia showed the highest correlation with heart failure, which is logical since the increase of potassium in the blood could lead to having severe heart arrhythmia, and this could result in heart failure. In addition, during the analysis of the number of people who had hyperkalemia and heart failure, it was found that 38.46% of patients suffer from heart failure, as well as hyperkalemia.

The following feature that was analyzed after blood potassium was the anion gap, which represents the gap or the space between the positive and negative charged electrolytes. As stated by (Anion Gap Blood Test: Medlineplus Medical Test, 2022), it is explained that the anion gap is considered necessary for maintaining blood with a balanced pH, and not having normal blood acidity could cause some health problems, including vomiting, shortness of breath, arrhythmia, and many other health problems. However, if it is left without treatment, it could cause more severe conditions, and the

health problems could worsen, and it might even lead to death. As a result, it is crucial to consider checking on anion gap while having regular blood tests to prevent the serious health conditions that could lead to death.

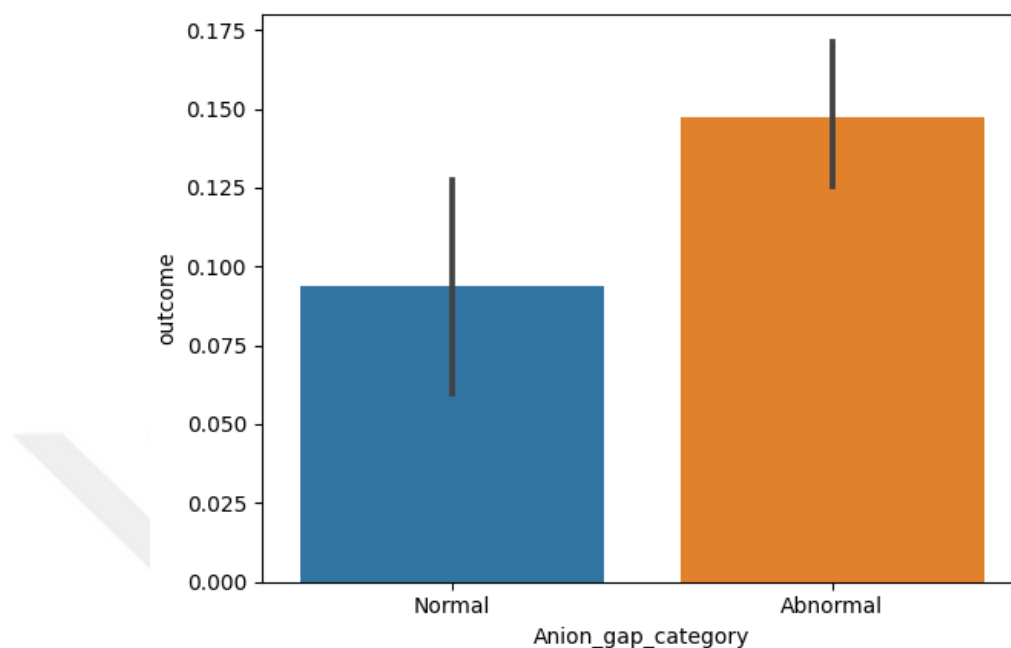


Figure 30 The Bar Chart of Anion Gap Vs Outcome

According to the figure above, which shows the analysis of the anion gap and the outcome, it is apparent that a group of patients had a high level of anion gap in their blood, which is considered an abnormal condition. The bar graph in the figure indicates two groups, normal and abnormal, and these groups contain patients with either normal or abnormal levels of anion gap. The first group, which is the one with the patients with normal anion gap levels, and that is with a range between 4 and 12 mmol/L. The second group is the one that contains patients with an abnormal anion level, which is more than 12 mmol/L, and this group showed a higher relation with heart failure. The results are acceptable since the changes and increase in anion gap could lead to having heart failure. It is crucial to mention that while analyzing the patients who had abnormal levels of anion gap, it appeared that 14.73% of these patients had both heart failure and high levels of anion gap. As a result, it is possible to end up having heart failure or even dead if the anion gap in the blood increases rapidly and no medical treatment is considered.

The last feature that was analyzed and visualized was Lactic Acid, which is considered as a kind of chemical the body produces while breaking down the carbohydrates in the body to get energy, as mentioned by (professional C. , Lactic Acidosis: Symptoms, Causes, Treatment & What It Is, 2024). Mostly, lactic acids are found in all body tissues, but red blood cells and muscles could produce more lactic acids than other body tissues. In addition, it is essential to note that normal lactic acid levels in the body should not exceed 2mmol/L since the kidneys and liver usually get rid of the excessive lactic acid. However, in some cases, people might have higher lactic acid level in the patient's blood, which would lead to some health conditions like nausea, muscle cramps, or even fever, and in some cases, it might lead to severe health problems like liver failure or even heart diseases. It is crucial to keep in mind checking the levels of lactic acid, and it is also important to know that not all cases of high levels of lactic acid cause health problems since some of the cases might have a temporary increase of lactic acid levels, which could be a result of exercising.

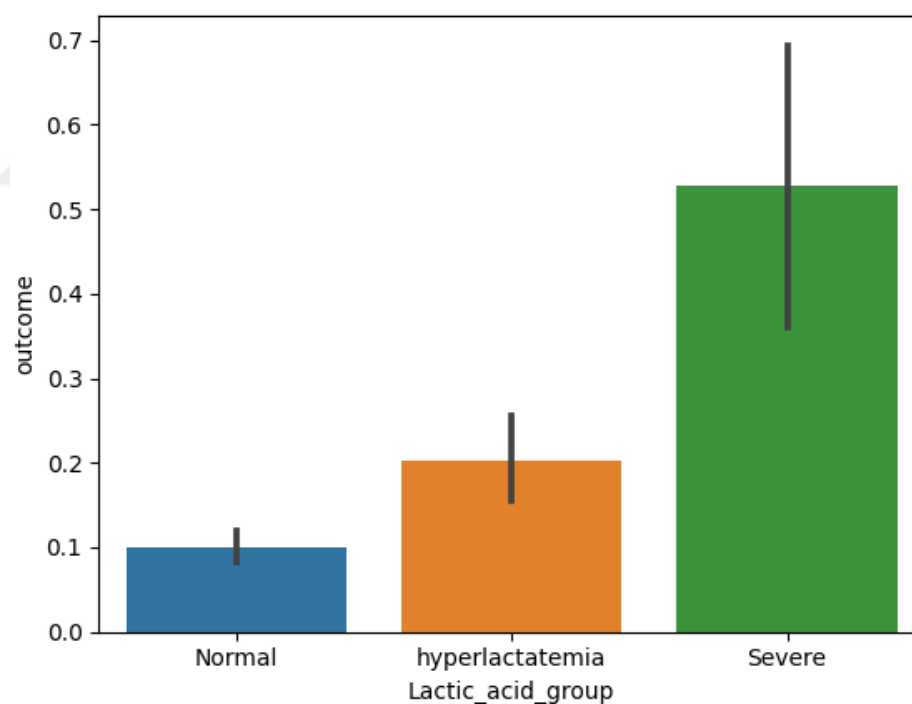


Figure 31 The Bar Chart of Lactic Acid Vs Outcome

Based on the figure that shows the relation between heart failure and the changes in lactic acids, it is obvious that most of the patients suffered from an abnormal level of lactic acids in their blood. The bar chart in the figure contains three groups, and each group presents a specific kind of patient. The first group, which represents the patients with normal levels of lactic acid, that is, as mentioned before, should not

exceed 2 mmol/L. The second group includes the patients with slightly higher levels of lactic acid, which is called hyperlactatemia, and this group of patients has lactic acid levels between 2 and 4 mmol/L. The last group, which contains all the patients with lactic acid levels higher than 4 mmol/L, and this is considered as a severe condition since this high level could cause all health conditions mentioned above. Additionally, it is clear from the analysis that the massive increase in lactic acid showed the highest correlation with heart failure among the three groups, and this kind of result is considered to be reasonable. The severe increase in lactic acid levels could cause severe health conditions, including heart failure, and this was maintained from the analysis since the results showed that 52.77% of patients with severe lactic acid had the highest risk of having heart failure.

After finishing the first step of analysis for the first dataset, the second dataset was also applied to the first step of analysis.

3.3.3.2 Data Analysis and Visualization for the second Dataset

This part is the continuous part of analysis, as it was done to the first dataset, this part includes data analysis and visualization for the second dataset. This dataset includes 12 different features together with 918 record of patients. Each feature was analyzed and visualized along with checking their correlation with the outcome.

At first, it was important to check the number of patients that ended up having heart failure in the dataset to see the exact percentage of heart failure patients since it was obvious from the histogram applied to the dataset that the amount of patients that were diagnosed with heart failure in this dataset was higher than the number of patients who were not diagnosed with heart failure. As a result, and to get a better idea about the percentage of patients with heart failure, a pie chart was applied to the outcome.

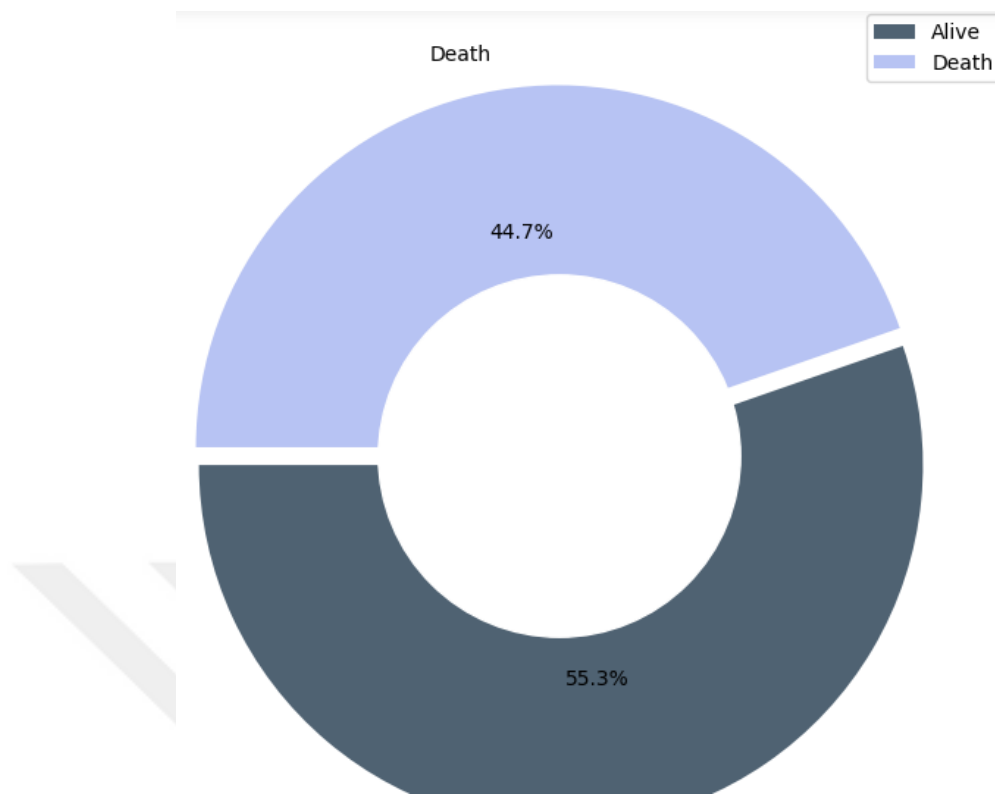


Figure 32 The Pie Chart of Target for Dataset 2

As it is shown in the figure above, a pie chart was applied to the outcome of the dataset to get an accurate percentage of people who were diagnosed with heart failure. Although the histogram showed that the patients with heart failure were more than the patients without heart failure, but the pie chart showed that the percentage of people who diagnosed with heart failure was only 44.7%. On the other hand, the percentage of patients who did not have heart failure is 55.3% of the total outcome, which means that still, the survivals of heart failure, even in this dataset, were more than the people who had heart failure. As a result, this pie chart shows how it was crucial to check the exact percentage of people who were diagnosed with heart failure to have a clear and accurate idea about the outcome of the dataset.

Furthermore, it was essential to know the percentage of the males and females in this dataset as well because this dataset, unlike the first dataset, showed some correlation between the gender feature and the outcome. As a result of that, it was important to know the overall percentage of males and females. Similar to the first

dataset, a pie chart was applied to the gender feature of this dataset to get clear idea about the number of male and female patients.

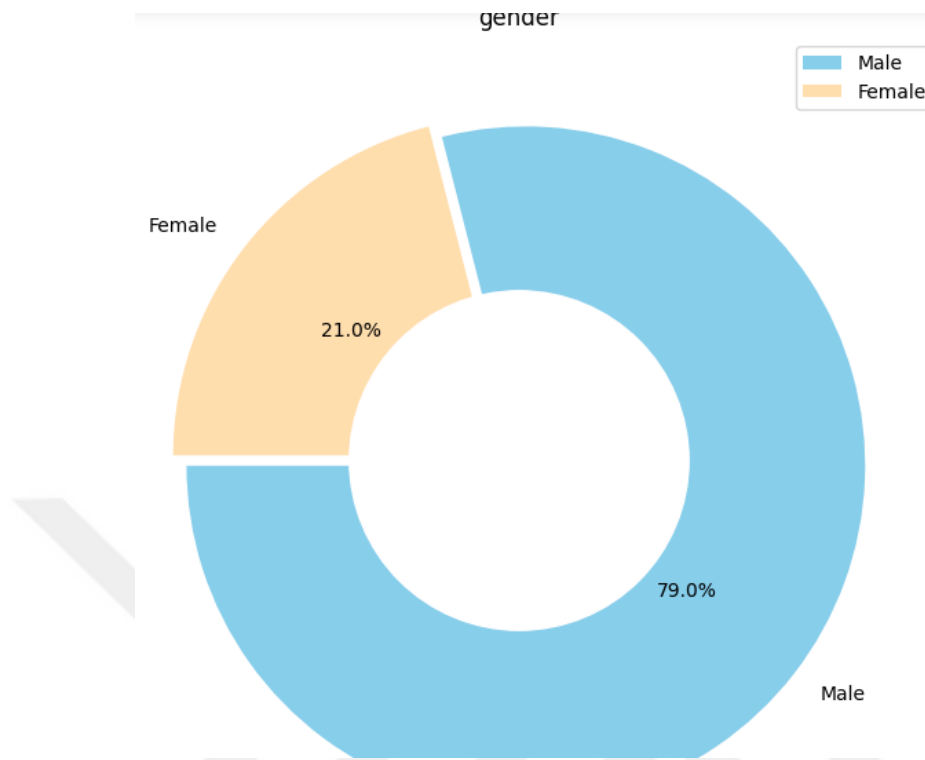


Figure 33 The Pie Chart of Gender for Dataset 2

According to the pie chart provided in the above figure that presents the percentage of male and female patients in the dataset, unlike the first dataset that had almost equal numbers of males and females, this dataset contains almost 80% of male patients and only 21% of female patients. This might be a sign that the number of patients that diagnosed with heart failure could be mainly from male patients.

The next feature that was analyzed and visualized for this dataset is Exercise Angina, which can be named as exercise-include angina pectoris. Angina is a kind of pain that affects the chest area, and it is a result of a lack of blood flow that is returned to the heart. Normally, angina comes in a picture of heaviness or pressure on the chest area, and it might even come as pressure on the arm or the jaw. As stated by , angina might occur when doing something that needs more blood in the body, which causes the heart to pump blood more complicated than usual, and when this happens, the heart needs to have blood that contains more oxygen to be able to pump it to the body. However, angina happens if the heart cannot get enough blood that is rich in oxygen. It is crucial to mention that there are multiple of angina types that could occur as a

result of different causes, but the one included in this dataset and was studied is the one that is caused by exercising.

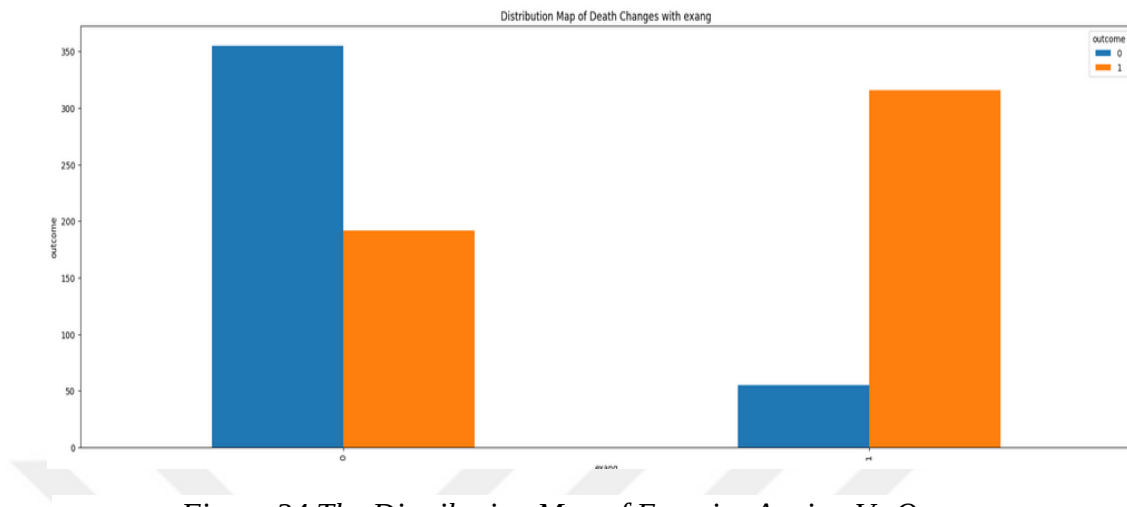


Figure 34 The Distribution Map of Exercise Angina Vs Outcome

According to the figure above, which presents the distribution map of the exercise angina with respect to the outcome, it is clear that the exercise angina is a binary feature that only includes two values, zero and one. Because of that, the feature was divided into two groups, and each of these groups contained a specific group of patients and contained two bars. The bars represent the patients who were diagnosed with heart failure, which is shown in the graph as an orange bar and patients who did not have heart failure, which is shown in the graph as the blue bar. The first group contained the patients who did not suffer from exercise angina, which contained the highest percentage of patients who did not have heart failure. On the other hand, the second group, which has the patients who suffered from exercise angina, showed the highest result of patients which diagnosed with heart failure. Although exercise angina does not apply long-term damage to the heart, but from the analysis above, it is clear that most of the patients face the risk of having heart failure while having exercise angina. In addition, it might be logical since angina is related to the heart in a way that angina only happens when the heart does not get enough blood that is rich in oxygen and eventually this could cause a damage to the heart leading to having heart failure.

The following feature that was analyzed in this dataset is the chest pain feature, which contains four types of chest pain. These types are examples of the kinds of pains that can affect the chest, and these pains are distributed from the slight pain to the heaviest pain. The first kind recorded in this dataset is the typical angina, which is a pain that comes with heaviness or pressure on the chest. The second kind of chest pain is Atypical angina, which is described as a pain that comes like stabbing, and it might increase according to some specific body positions. It is essential to mention that this type of chest pain might come and go, so it is not continuous. The third contained in the chest pain feature is Non-Anginal pain, which is also a pain that is felt in the chest area but does not indicate a problem with the heart. The last type was recorded in the dataset is Asymptomatic, which considered as the most critical type since it does not show any symptoms but causes great damages in the body, specifically the heart, since by having asymptomatic, the heart works harder than usual and does not get enough oxygen in the blood, which could lead to a severe condition like silently having heart failure.

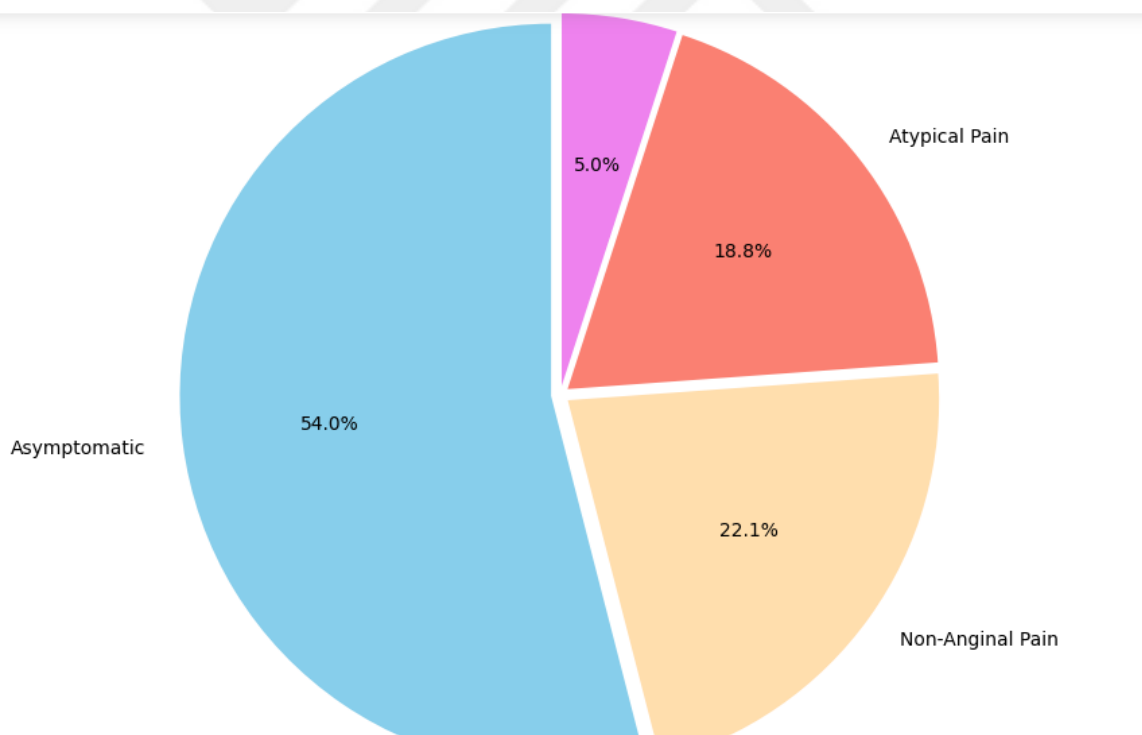


Figure 35 The Pie Chart for the Types of Chest Pain

According to the pie chart above, which shows the percentage of patients with each type of chest pain. It is obvious that most of the patients had the Asymptomatic

type, which was 54% of the patients in the dataset, and this is ideal since this dataset is considered to study the patients who are at risk of heart failure. The second type that included 22.1% of the patients in the dataset is Non-Anginal pain, which is considered the second critical chest pain type among the other types. The other two types, which are Typical and Atypical, did not contain many patients, and they only had 5% and 18%, respectively. Moreover, after having a clear idea of the chest pain types and how many patients recorded for each type, an analysis was made to the chest pain feature to see how it is related to the heart failure.

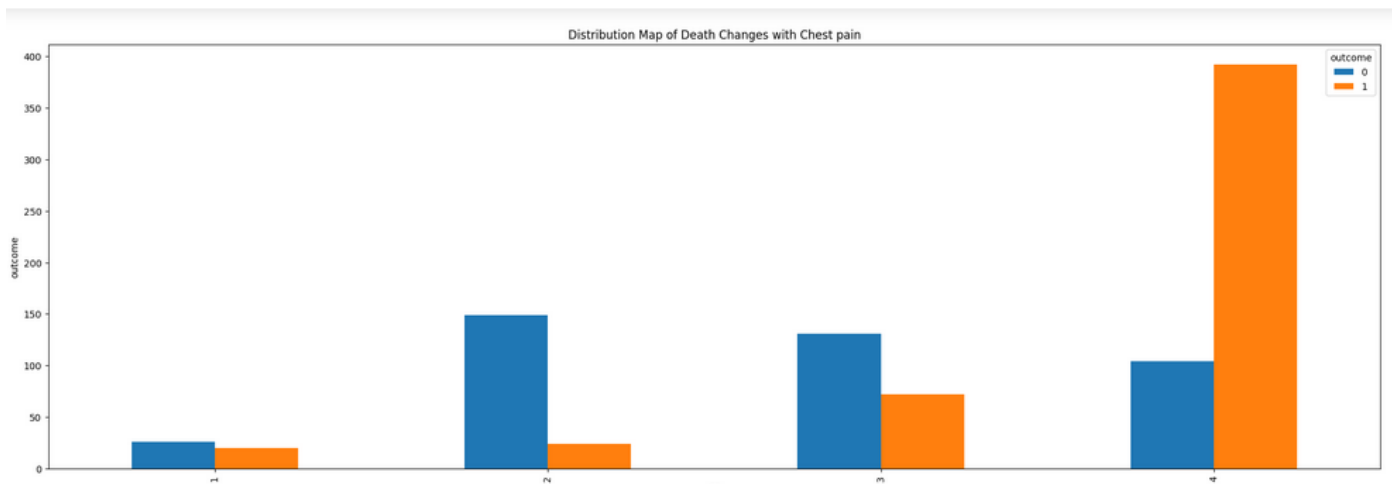


Figure 36 The Distribution Map of Chest Pain Types Vs Outcome

Regarding the above figure, which contains the distribution map of chest pain features with respect to the outcome. The map shows how much each type is related to heart failure, and it is clear from the first group, which contains the patients that have the Typical angina type, that this type almost did not show any relation with heart failure, neither on the survivals of heart failure nor the patients with heart failure. This might be a result of the lack number of patients who suffer from this type of chest pain. The second group in the map, which is the group that includes patients who suffer from Atypical angina, showed some results that most patients with this chest pain type survived heart failure. The third group of chest pain, which included people with Non-Anginal pain type, showed some significant results, as it showed more patients with heart failure and fewer patients who survived heart failure compared to the second group. The last group in the mad, which includes people who suffer from Asymptomatic, showed the highest results of correlation with heart failure and the lowest survivals of heart failure when compared to the second and third groups. As

mentioned before, it is logical to get such results since Asymptomatic is considered as one of the critical types of chest pain, which affects the heart in a serious way. The issue with this type is not giving any signs to the patient that they suffer from Asymptomatic, and since this type is directly related to the heart, the continuous attack of this chest pain type without knowing that it is attacking the patient's chest could cause severe damages to the patient, especially to the patient's heart.

The following analysis was applied to the gender feature, since the type of gender showed some correlation with heart failure, it was essential to study what is the relation between heart failure and gender, also to see which type would show more correlation with heart failure.

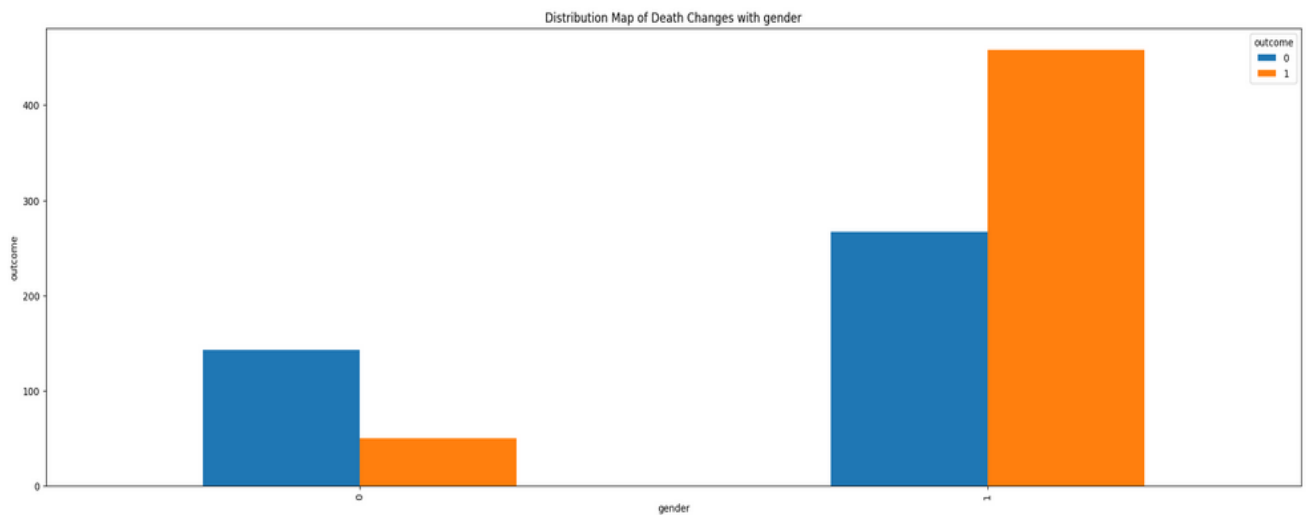


Figure 37 The Distribution Map of Gender Vs Outcome

The figure above shows two groups and each group has two bars. The groups represent the gender feature, which contains males and females, and the bars represent the patients who survived heart failure and the ones who were diagnosed with heart failure. Moreover, as it is clear from the graph that the first group, which refers to the females in the dataset, did not show much relation with heart failure; in fact, most of the females showed a survival of heart failure. On the other hand, males showed a significant correlation with heart failure, and most of the male patients were diagnosed with heart failure. These results might not be so accurate since when analyzing the gender feature, it showed that most of the patients in this dataset were males, and this might be the reason behind these results; after all, from what is given in this dataset that the males have more risk of developing heart failure than females.

The next feature that was analyzed and visualized was the age feature, which represents the age range of 918 of patients in this dataset. The age range recorded for this dataset was between 28 and 77, which made it easy to read from a distribution map as a feature analyzed with respect of the target.

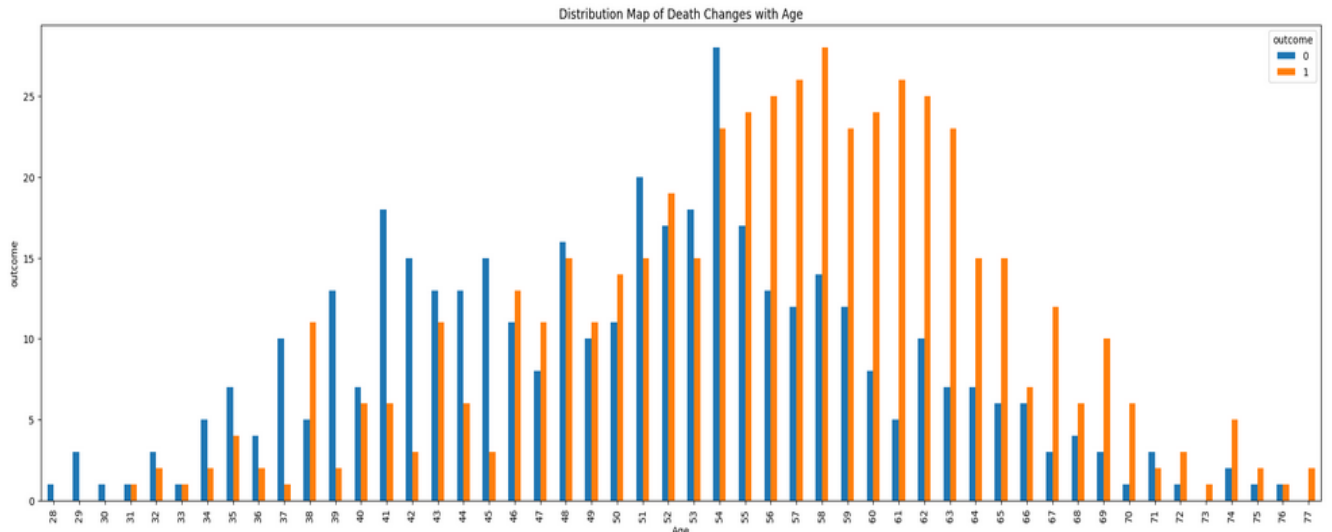


Figure 38 The Distribution Map of Age Vs Outcome

The figure above contains the distribution map of the age with respect to the outcome, and it is possible to read from this map the ages of the patients and their correlation with heart failure. As it can be seen that the patients with younger ages, like the ones who were 28 years old, did not contain any person with heart failure. However, the older the patients get, the more correlation to the heart failure they give, as starting from age 55, patients were getting more risk of having heart failure and did not show much survival at that point. In addition, it is clear that patients between 73 and 77 years old did not record any survival from heart failure. In order to get a clearer picture of these patients and their risk of having heart failure, another graph was plotted.

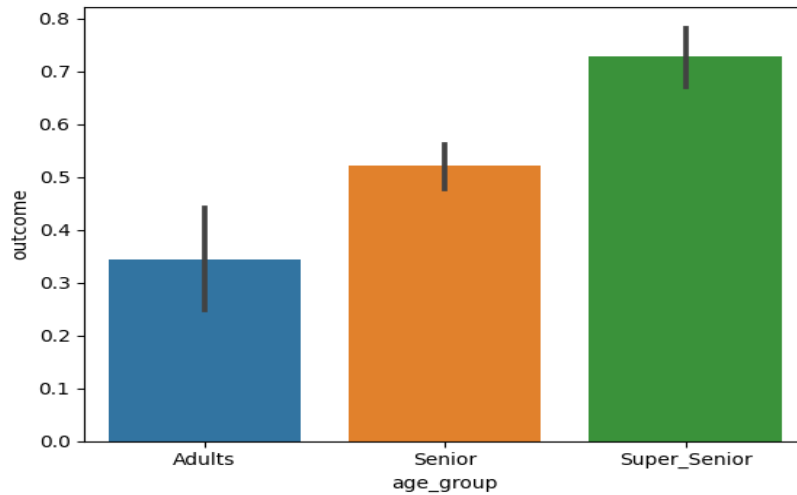


Figure 39 The Bar Chart of Age Vs Outcome

Based on the figure above, which presents the age groups and their relation with heart failure, it is clear that the patients were divided into three groups in order to get a more clear idea of which group of people would show more correlation with heart failure, in other words, which group would have more risk of having heart failure. In the beginning, the age feature was divided into three groups, which are Adults that present the patients ages between 20 and 39 years old. The second group was formed as the senior group, which included patients with ages ranging between 40 and 59 years old. Lastly, the third group was made to contain patients with ages between 60 and 80 years old, who had the name of super senior. Furthermore, it is possible to see that the super senior group showed the highest correlation with heart failure among the three groups, which indicates that people in this group have a higher risk of getting heart failure than the people in the other groups. This result is quite logical since it is possible to get weaker muscles while getting older, which includes the heart muscle, and is the reason for the importance of getting the heart checked every now and then.

Another feature that was analyzed and visualized in this dataset was fasting blood sugar. Fasting blood sugar is considered a kind of test that doctors make to see the sugar levels in a patient's blood. It is crucial to measure blood sugar since high levels of sugar in the blood could indicate a risk of having diabetes. The test for calculating fasting blood sugar is done when ensuring the patient is still fasting, and this is an important factor because anything the patient eats would increase the insulin in the blood, which would give false results of this test. As stated by (Diabetes Tests, 2023), it is essential to mention that there are specific levels that are considered a normal level of fasting blood sugar, and this level is recorded as any number equal to

or below 99mg/dL. However, if the level of fasting blood sugar was between 100 and 125mg/dL, this might indicate pre-diabetes, which is the first sign of having diabetes. If the level of the fasting blood sugar test is higher than that, it means that the person has diabetes.

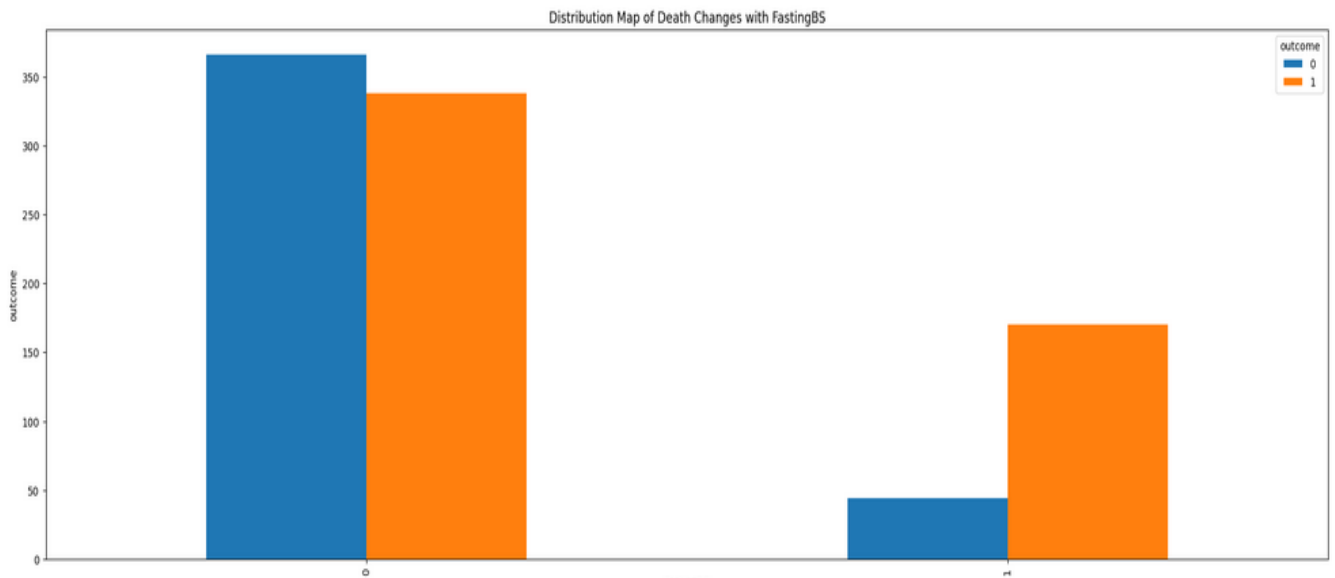


Figure 40 : The Distribution Map of Fasting Blood Sugar Vs Outcome

The distribution map plotted in the figure above shows the relation between fasting blood sugar and outcome, and as it is clear from the map, the fasting blood sugar had two outputs. The first group in the map presents the patients that had low fasting blood sugar results, and the second group included patients that had a fasting blood sugar higher than 120, which indicates that these patients suffer from diabetes. The graph, which contains the patients who do not have diabetes, showed that the first group had some significant results in relation to heart failure. It is clear that people who did not have diabetes were at risk of heart failure as well as not having that risk, but it is important to mention that people who had no diabetes had a higher chance of not having the risk of having heart failure.

In addition, this can be a result of having other features that might affect the fasting blood sugar results. On the other hand, it can be noticed that the second group showed that most of the patients who suffer from diabetes were at risk of having heart failure, which indicates that the increase in the fasting blood sugar, or having diabetes, would increase the chance of getting heart failure.

The next feature that was visualized is the resting blood pressure, which indicates whether the patient has blood pressure or not. Blood pressure indicates the measurement of how powerful the blood is flowing in the vanes, and it should not be more than 120/80 mmHg when the person is at rest. As mentioned on (professional C. , What is Blood Pressure?, 2024), high blood pressure can happen at any age, and it is a common health condition, but it is more common for specific people who have diabetes, who smoke, or even elder people whose ages pass 60 years old. It is also mentioned that blood pressure is considered as a severe health condition since, if it is not treated, it could cause life-threatening conditions such as kidneys failure, stroke, heart failure, and many other conditions that could lead to the loss of life.

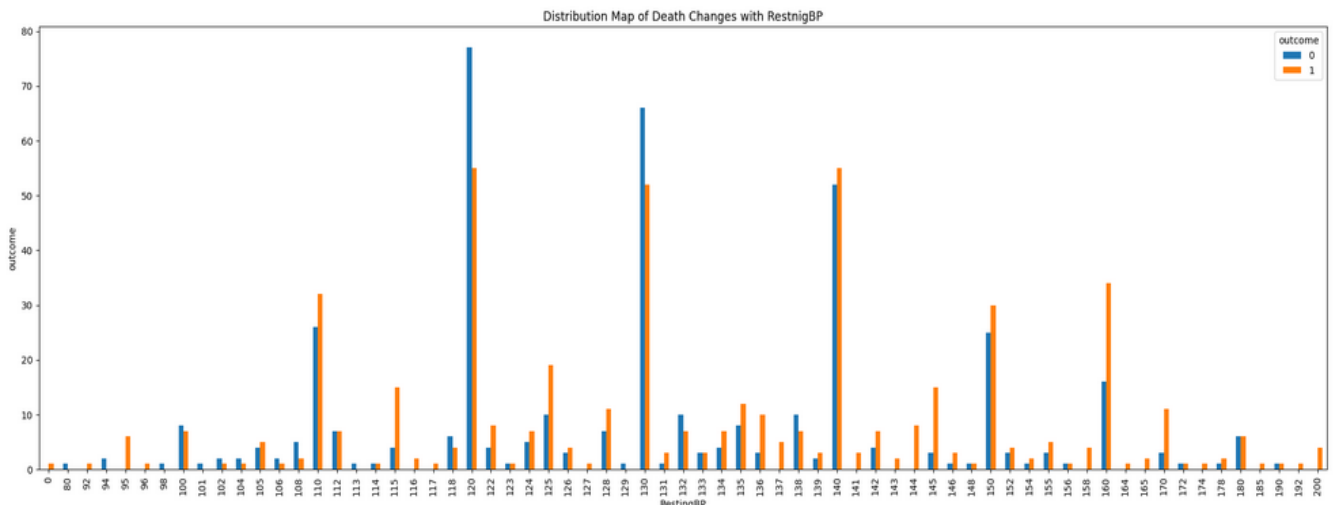


Figure 41 The Distribution Map of Resting Blood Pressure Vs Outcome

The distribution map, which is shown in the figure above, is a graph that includes the relation between resting blood pressure and heart failure. The graph shows that almost all the records of patients who had a resting blood pressure of more than 120 mmHg were at risk of having heart failure. It can be seen from the graph that after 120 mmHg, all the orange bars, which present patients with heart failure, showed higher results than the blue bars, which represent patients who did not have heart failure. Moreover, it is also visible that there were no surviving patients from the risk of having heart failure, especially in the last three bars, which indicates patients with 190 mmHg blood pressure and higher. In addition, to get a more clear idea of the relationship between resting blood pressure and heart failure, another graph was plotted.

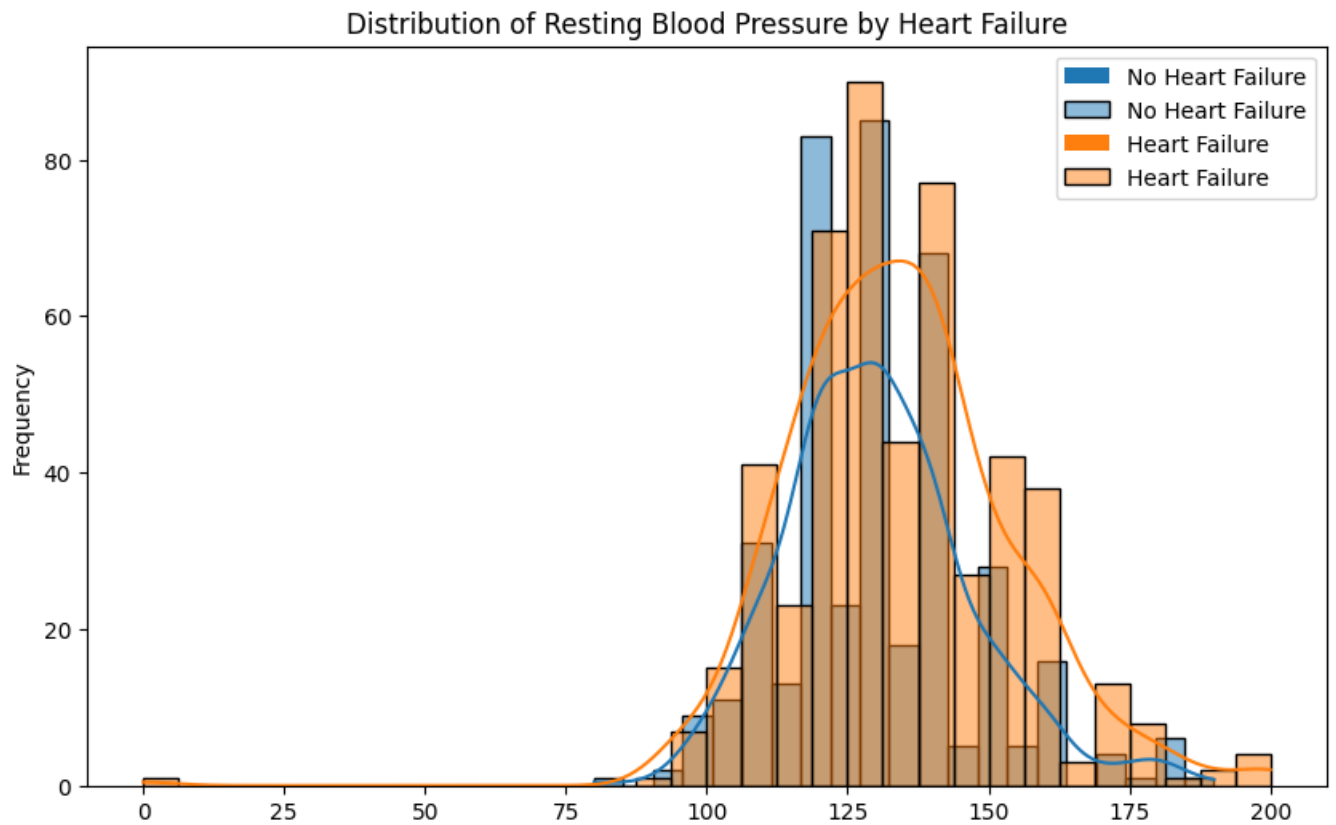


Figure 42 The Distribution Map of Resting Blood Pressure Vs Outcome

The above graph shows more clearly the results of patients who have high blood pressure and had heart failure. The map has four variables that present heart failure, two of which present patients with heart failure, which are colored in orange, and the other two present survivals of heart failure, which are colored in blue. It is clear from the distribution map that the orange line, which indicates patients with heart failure, showed higher results than the blue line, which indicates survival patients from heart failure. As a result, this graph explains exactly how high blood pressure affects people, that is the higher the blood pressure they have, the more risk of having heart failure they get. According to these results, it is crucial to control the increase in blood pressure; if the patient does not seek medical attention when having high blood pressure, it would affect the patient's health and might lead to severe health conditions like heart failure.

The last feature in this dataset that was analyzed and visualized is resting ECG. Resting ECG is a test that can be done on patients while resting and does not indicate any surgeries. This test detects the problems that can be happening in the body, like abnormalities, which include arrhythmia and other issues. The resting ECG test was

included in this dataset because it is the kind of test that doctors do to the patient to have an idea if the patient is having heart failure or not. Moreover, this feature was included in this dataset in, which has three values, and each value presents a specific level of resting ECG test. The first value, which included the patients with normal levels of the resting ECG test. The second group involved the patients who had some abnormalities with the T wave during the test reading. The last group, which included patients who already had problems and might even have ventricular hypertrophy. In addition, to get an understandable idea of this test and its correlation with heart failure, a bar graph was plotted.

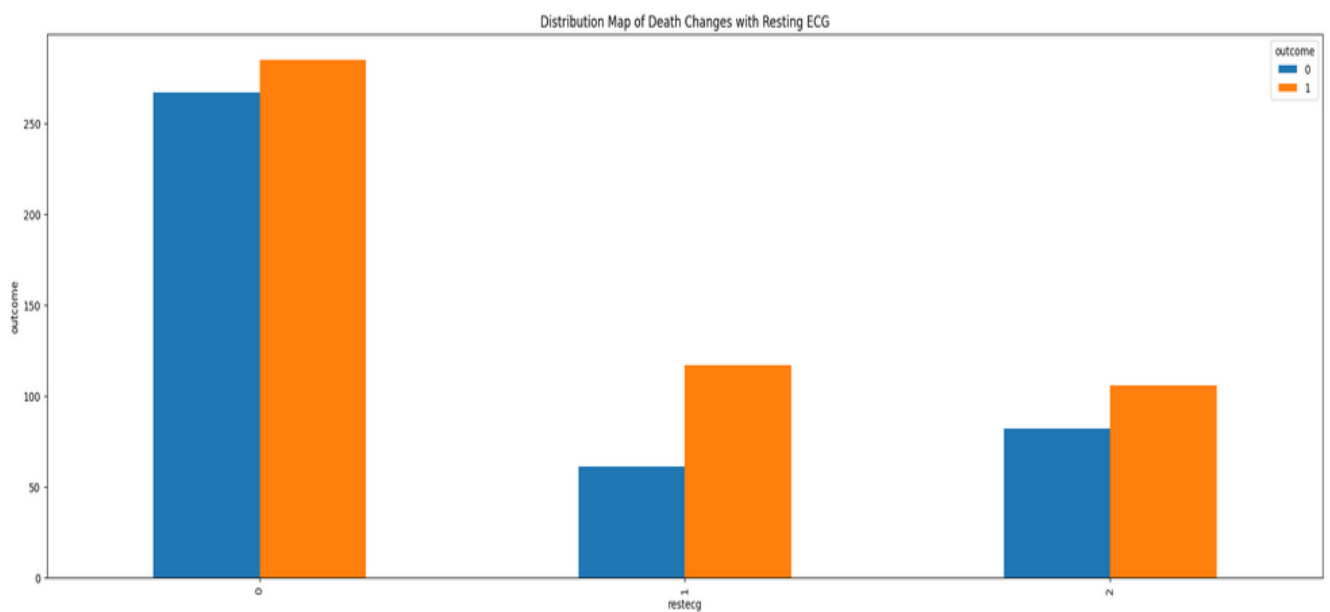


Figure 43 The Distribution Map of Resting ECG Vs Outcome

The figure above shows the relation between the values of resting ECG and the outcome, and it is showing an impressive results. The graph included three groups as each group represents one of the resting ECG test values, and each group is showing a relation with heart failure, which is represented by the blue and orange bars. The blue bar here indicates the patients who did not show results of having heart failure; on the other hand, the orange bar represents the patients who had heart failure. It is impressive to see that the three groups showed a significant relationship with the heart failure feature, in which that the first group showed the highest correlation with heart failure; at the same time, it showed that the highest result of people who did not have heart failure. However, this result could be affected by other features that would show a relation with heart failure, meaning that if the patients with normal rest ECG test but

had high blood pressure, as an example, this might affect the results of the comparison in the graph. The second group showed the lowest result of patients who did not have heart failure, and this might also be effected by other features, because this group should have showed more survived patient from heart failure than the last group, according to medical professionals. As a result of this analysis, it is always important to maintain normal ECG and take all the necessary treatments if the resting ECG test shows any kind of abnormalities.

3.3.3.3 Data Modeling and Prediction for both Datasets

After finishing the first step in the analysis, which was analyzing and visualizing each feature on its own and checking its correlation with the outcome, the second step was applied. This step includes the splitting method, which is one of the most crucial steps that should be used when building machine-learning models. Moreover, the splitting method works by splitting the dataset into two groups, and each group represents part of building the machine learning models. The first group is the training group, which will contain all the features that will be used to train the models. On the other hand, the second group is the test group, which will include the features that will be used to test the machine learning model's accuracies. These two groups are used together to ensure that there would not be a high risk of over-fitting since the over-fitting happens when the dataset is very complex in a way that makes the training data over-fit, and as a result of that, the testing part would perform poorly. As a result, the splitting method is considered a crucial method while building the dataset model to ensure that these models would give better performance and reliable accuracies. It should be mentioned that this method was applied separately for each dataset in this project.

The first part of splitting data is to define the variables x and y , which are part of the splitting data technique and are representatives of the features in the datasets. As for x , it represents the matrix for each feature in both datasets since each column in the datasets is considered a feature, and their rows are the records for each feature. On the other hand, y represents the target for each dataset, which is presented by the feature outcome for both datasets. In addition, it is essential in the splitting method to assign a specific percentage for the test set, which was made for both datasets. In this case, both datasets were assigned 30% of the data that was used to test the models, and the other 70% of the data was used for training. Furthermore, after finishing the

splitting data method, it was important to equate the train-test function with four specific variables. The first variable is a representative of the feature matrix that includes a subset, which is made of rows from the main feature matrix in the x variable, and this subset that is included in the first variable is the x-train.

The second variable, the x-test, is a representative of the feature matrix that includes the left row from the main feature matrix in variable x, and the x-test variable only uses the left rows from x without including the ones from the training set. The third variable is called y-train, which represents the target variable that is connected to the training set, and it contains a subset that is taken from the main target variable y, which is also connected to the rows in the x-train. Lastly, the last variable that was equated to the train-test function is the y-test, which is a representative of the target variable with the test set, and it contains the left values from the subset that was taken from the original target variable y that is connected to the rows in the x-test variable.

The following step was to use the standard scaler method, which is a technique used to standardize the features included in the splitting method by deleting their mean and scaling them. The standard scaler method was used with the fit-transform method; the standard scaler method here calculates all the means and standard deviations for each feature included in the x-train variable, and then, the fit-transform technique returns all the transformed training data into a specific variable, which called x-trainS, which will contain features that has their mean almost equal to zero and their standard deviation equals to one. Moreover, after using the standard scaler method to calculate the means and standard deviation to the training data in x-train and the fit-transform to return the transformed training data, the same method was applied to the test data in a way that the transformation calculated on the training data x-train to the x-test. After that, the transform method was used from the standard scaler to standardize the test data using the mean and standard deviation, which was calculated from the training data without recalculating it; then, it returns all the transformed test data into a specific variable, which called x-testS. As a result, the techniques used above are all made to ensure that the training and testing data are applied to the same method. Therefore, they will be scaled in the same way.

In addition, before starting to build the machine learning models, three lists were created to ensure that the accuracies from all the models were collected under one list. There were three lists because the accuracies were gathered from three

different techniques, and each technique has its own effect on the machine learning models' accuracies. The first list was made to collect all the accuracies that were calculated by the machine learning models with no use of any other techniques, and this list was created by giving it the name of `acclist`, which indicates the accuracy list, and then was equaled with an empty square brackets. The second list was created in the same way, but it was made to the machine learning models that they were applied to the cross-validation technique, which gave different accuracies from all the models. The last list was made the same way the first two lists were made, and this one included the accuracies that were given from the machine learning models but after they were applied to the principal component analysis technique, which also gave different accuracies from the first two lists.

The next step in the analysis was building the necessary machine learning algorithms to get the accurate predictions of these models. The models used in this project were applied to both datasets, which were KNN, Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine, Ada-Boost, Extreme Gradient Boosting, Random Forest, Extra-Trees, Artificial Neural Networks, Linear Discriminant Analysis, Cat Boost, and LightGBM. All the algorithms were applied using the same steps mentioned below.

First of all, the first model was built for this project is the K-nearest neighbor, which was identified as `knn` in the project. The model was imported from a library called `sklearn.neighbors`, and this library allows users of Python to build `knn`. After importing the library and creating `knn`, it was important to consider the number of neighbors that were needed in the project while creating the class `knn`, and at this point, a three neighbors were specified in the `knn` class to consider while making a prediction. Furthermore, the model was used to be fitted to the training data, which was done by using the `fit` method. The `fit` method allows the model to be trained by learning patterns between features in the variable `x-train` and the target values in the `y-train`. This method is used to train the model and make it learn the patterns from the features in the `x-train` and the target from the `y-train` in order to be able to start making predictions. Lastly, the `score` method was applied to the `knn` model to calculate the accuracy of the model on the test sets `x-test` and `y-test`. The `score` method allowed the model to calculate the accuracy and put it under the variable `knn`, and from there, this result was assigned to the accuracy list for both datasets.

The second model that was built for this project is Naive Bayes, which had the name gnb in the project. To build the Naive Bayes model, it was important to follow the same steps that were applied to build the KNN model, which started by importing the GassuianNB library. The thing about Naive Bayes is that there were not any neighbors or classes to consider while building the method, so it was used as it was, and nothing was included during the building process. Then, the model was also fitted to the training data, which included in the x-train variable and y-train variable. The fitting allowed the Naive Bayes model to learn the patterns included in the variable x-train and the target y-train. Moreover, the last step was applied, which is using the score method that allowed the model gbn to calculate its accuracy on the test sets x-test and y-test. As a result, the accuracy of the model was saved under gnb, and from there it was assigned to be included under the accuracy list for both datasets.

The third model that was built for this project is the Decision Tree, which had the name dtree in the project. To build the Decision Tree model, there were some extra steps when compared to the building of the other first two models. When importing the model, it was crucial to specify a couple of things to ensure the model would work properly. The first thing to do was to include a parameter from criterion called entropy, which is a kind of parameter that is used to split the nodes in the Decision Tree model. Moreover, choosing the entropy for this model was to ensure that the model would use the information that is calculated by entropy to find the best split for each node in the model. In addition, the depth for the model was added, which presents the longest path the tree can have from its root to the last node. The path, in this case, was added as 2, which means that the model will only have two levels of depth. This choice would enhance the accuracy of the model since choosing a specific depth would help preventing the over-fitting. After that, the model was applied to the fitting method as well, which was done by fitting the model to the training data that included the variables x-train and y-train.

Lastly, the score method was applied, which it allowed the model dtree to calculate its own accuracy on the test variables x-test and y-test. As a result, the accuracy of the model was obtained under dtree and then was included in the accuracy list.

The next model built in this project was Logistic Regression, which is given the name lr for both datasets. The model was built in a way that needed some different

steps from the previous models. At first, when importing the model, it did not need any specific added steps as it was in the Decision Tree model; it only needed to import the necessary library for Logistic Regression to start building the model. After importing the model, the fitting method was applied, and that was by fitting the model to learn from the training data that included the features from the x-train variable and the target from the y-train variable. Then, a prediction variable was made, which included the prediction of the model lr that was generated by the test data included in the x-test. Lastly, a prediction function was generated to compute the prediction accuracy that was made by comparing the with the values that were found true in the y-test variable, in which that the variable of prediction compares the results it has to the true values from the y-test variable and returning the correct prediction to the label. At the end, the prediction of the Logistic Regression model was also included in the accuracy list for both datasets.

The model was built and used in this project after the Logistic Regression model was the Support Vector Machine, which was given the name sv in this project. The Support Vector machine was built almost the same way that the previous models were built, which needed its library to be imported into the notebook. The model was applied to the fit method as the other models were to be fitted to learn from the training sets, which included the features of the x-train variable and the target from the y-train variable. After that, the model was applied to the score method, which allowed the model to calculate its accuracy on the test variables that included the x-test left features and the y-test target. Furthermore, after getting the accuracy from the Support Vector Machine model, it was contained in the accuracy list for both datasets.

The next model that was structured in this project is the Ada-Boost model, which gave the name adab as a shortened name for the model. To start structuring the model, the library of Ada-Boost was imported into the notebook. After giving it the shortened name adab, the model was applied to the fitting method, which was used to fit the model in order to learn from the training sets, which are the features from the variable x-train and the target from the y-train variable. Moreover, the model, as the other models, was applied to the score method to be able to get its accuracy, and that is possible from the score method since this method allows the model to calculate the accuracy of the test variables that contain the left features from the x-test variable and the target in the y-test variable. In the end, the accuracy of the Ada-Boost model, which

is saved in the `adab` variable, is also included in the accuracy list, which is contained for both datasets.

Another model was applied in this project to get its accuracy, which is the Extreme Gradient Boosting. The model was given a shortened name in this project, which is `xgb`. The model needed to be imported in the notebook in order to have the ability to build it, which was imported and assigned under the `xgb`. After importing the necessary library, the model was applied to the fitting method to fit the model on the training sets that contain the features from the `x-train` variable and the target from the `y-train` variable. Moreover, it was also applied as the other models to the score method to be able to get the model's accuracy, which is reasonable since the method gives the ability to the model to calculate its accuracy on the test variables that include the rest of the features in the `x-test` variable, which are left from the training variable, and the target in the `y-test` variable. In addition, the result of the Extreme Gradient Boosting model accuracy was also included in the accuracy list that was made for the two datasets.

The following method was built in this project is the Random Forest model, which gives the name `rf` as a shortened name. The model had to be imported from a library, which is `sklearn.ensemble`, and the library was called into the notebook to be able to build the model. After importing the necessary library, the Random Forest model was applied to the fitting method, which indicated the fitting of the model on the training sets that had the features from the variable `x-train` and the target from the `y-train` variable. Furthermore, it was applied as well to the score method like the other models, and this method was necessary to get the accuracy of the model. The score method allowed the Random Forest model to calculate its accuracy on the test variables, which contain the variable `x-test` that included the left features and the variable `y-test` that included the target feature. After forming and applying the model to the training and test variables, the accuracy of the model was assigned to the variable `rf`, and that variable was included in the accuracy list in the two datasets.

The next machine-learning model that was built to be used for this project is Extra-Trees, which was given the name `extrat` as shortened name to ease the use of the model. To have the ability to build the Extra-Trees model, it was essential to import it from the library `sklearn.ensemble`. After importing the necessary `ExtraTreeClassifier`, the fitting method was applied to the model to be able to build it, which was by fitting

the model on the training sets that included the features from the x-train variable and the target from the y-train variable. In addition, to be able to finish building the model and get its accuracy, the model was applied to the score method like the other models. The score method gave the ability to the Extra-Trees model to compute its accuracy on the test sets, which included the remaining features in the variable x-test and the target in the y-test variable. As a result, the model was formed and applied to the necessary method to get its accuracy, which was assigned under the name `extrat`, and from then, it was included in the accuracy list like the other machine learning models.

The following machine learning model was used in this project is the Artificial Neural Networks, which was assigned with the name `ann` to be used in this project. The Artificial Neural Networks model is not like the other models, and it needed some specific libraries and some extra steps to be able to build it. The building process of the Artificial Neural Networks started by importing a specific library, which is the tensorflow library, and from this library, the keras module was imported. The keras module is an important module for building the neural networks model as it is the interface to train and build neural networks. The next thing to import is the Sequential class, which is imported from the `tensorflow.keras.models` module; this class is considered as a linear amount of layers that makes it easy to add layers to the neural networks model. In addition, the last thing to import is the Dense layer class, which is also imported from the `tensorflow.keras.models` module and this layer is considered as a standard totally connected layer that makes the neuron in the prediction layer connect to each neuron in the layer. The following step to build the Artificial Neural Networks was to create a new sequential neural network model, and then the dense layers were added. The first dense layer was fixed to include 64 neurons and was given the input from the features in the x-train variable. It was applied to the activating function, which is the Rectified Linear Unit. At this step, the model created a layer, which took the input from the x-train variable and was applied to the linear transformation that was followed by the Rectified Linear Unit.

The second dense layer was created to include 32 neurons, which took the output of the first layer and used it as its input, and then it was applied again to the linear transformation. The last layer of dense, which included only one layer, was made to present the output and it contained only one unit. Moreover, to compile the model, the optimizer, the function loss, and the metrics were assigned, which was, first, by

using the adam optimizer, which is a way that machine learning uses to train the neural networks model. The other step was assigning the function loss, which was contained as the binary crossentropy and that was to calculate the prediction from the probabilities of the binary classification problems and the true labels.

The last parameter was assigned is the metrics, and since this project is about classification and finding accuracy, the metrics contained the accuracy function, which was included to assess the model while the model is trained and tested, and it helps to calculate the amount of classified sample. In addition, after building the model, the fitting method was applied to the models, which helped the model to be fitted to the training sets that included the features from the x-train variable and the target from the y-train variable. In the end, the accuracy of the model was contained in the parameter ann, which then was added to the accuracy list like the other methods.

The next model built for this project was the Linear Discriminant Analysis, which was given a shortened name lda. To create this model, only one library was included, which is `sklearn.discriminant_analysis` that was used to import the Linear Discriminant Analysis model. The model used the fitting method as well, to be able to build the model, which included fitting the model in the training sets that contained the features from the x-train variable and the target from the y-train variable. In addition, to get the model's accuracy, it was applied to the score method, which gives the model the ability to calculate the accuracy on the test sets that include the left features in the x-text variable and the target feature from the y-test variable. In the end, after the model had its accuracy under the parameter lda, the accuracy was added to the accuracy list to be compared to the other models' accuracies.

The following machine learning model that was built and used in this project is Cat Boost, which was given the shortened name cb. To start building the Cat Boost model, it was important to import it from the Catboost library. Also, the Pool class was imported from the library. The Pool class is a critical factor in building the Cat Boost model, in which it allows using the features and labels in an efficient way. After importing the necessary classes from the library, the Pool class was built by dividing it into two objects. The first Pool object, which included the training set for the Pool class, was built by including the x-train variable, which contained the features of the training data and the labels connected to the training set in the y-train variable. In addition, the other Pool object included the testing set, which consisted of the

remaining features from the x-test variable and labels connected to the y-test variable. The aim of preparing the Pool objects was to ensure that the data were trained and tested with the Cat Boost model. In addition, after finishing creating the parameter cb and the Pool objects, the fitting method was applied to the model. However, the fitting method here was applied differently from the way it was applied to the other models; since Cat Boost relies on the Pool object, it used the fitting method by using the train-pool, which is the Pool object that includes the training sets, and used the test-pool, which is to evaluate the performance of the model on the test sets. After that, the accuracy was maintained under the parameter cb and was added to the accuracy list, which included all the accuracies of the previous models.

The last machine-learning mode used in this project is LightGBM, which is given the shortened name lgb. To build the model, some methods were applied in steps, and the first step was to prepare the data to be fitted for the model. Firstly, the data was divided into training data and testing data, and each of these groups included specific parameters. The training data group it included the features from the x-train variable and included the labels that are connected to the target in the y-train variable. For the testing group, it included the remaining features from the variable x-test and the target in the variable y-test. However, the test group contained an essential object, which is reference and it was assigned to be the training group, and this is a step that should be done when the data contains different features. After that, the parameters of the model were assigned to enhance the performance of the model, and that step started by including the objective, which was assigned as a binary, and that is to indicate the binary classification. The following parameter was included is the metric, which included the binary-error and that means that it will measure the rate of errors for the binary classification. The third parameter is the boosting-type, which indicates the type of boosting algorithm that can be use and in this project the Gradient Boosting Decision Tree was assigned to be used as a boosting algorithm. The fourth parameter included is the num-leaves, which indicates the number of leaves that can be applied in each tree in the Gradient Boosting Decision Tree, and it was given the number 31 leaves for this case. The next parameter is the learning-rate, which is referred to the number of adapts the model can do during the training, and it was given the rate number 0.05. The other parameter assigned to model the LightGBM model is the feature-fraction, which is referred to the feature fractions that should be considered

while building the trees, and this parameter helps the model control over-fitting. The last parameter that was included in modeling LightGBM is the early-stopping-rounds, which indicates the number of times that the model should be considered when the validation set performance does not improve, and this number was 10 for this project.

After the previous step, the number of rounds was given to the model, which refers to the number of times that a new decision tree is added. Moreover, the model was applied to the training data, the parameters assigned before, and the testing data to calculate its accuracy. In the end, the accuracy of the LightGBM model was calculated and saved under the parameter lgb, which was then included in the accuracy list to be compared to the other accuracies from the other models.

Additionally, after collecting all the models in one list to make a comparison between the models' accuracies, a bar graph was plotted to get a clear idea on which of the models gave the highest accuracy prediction.

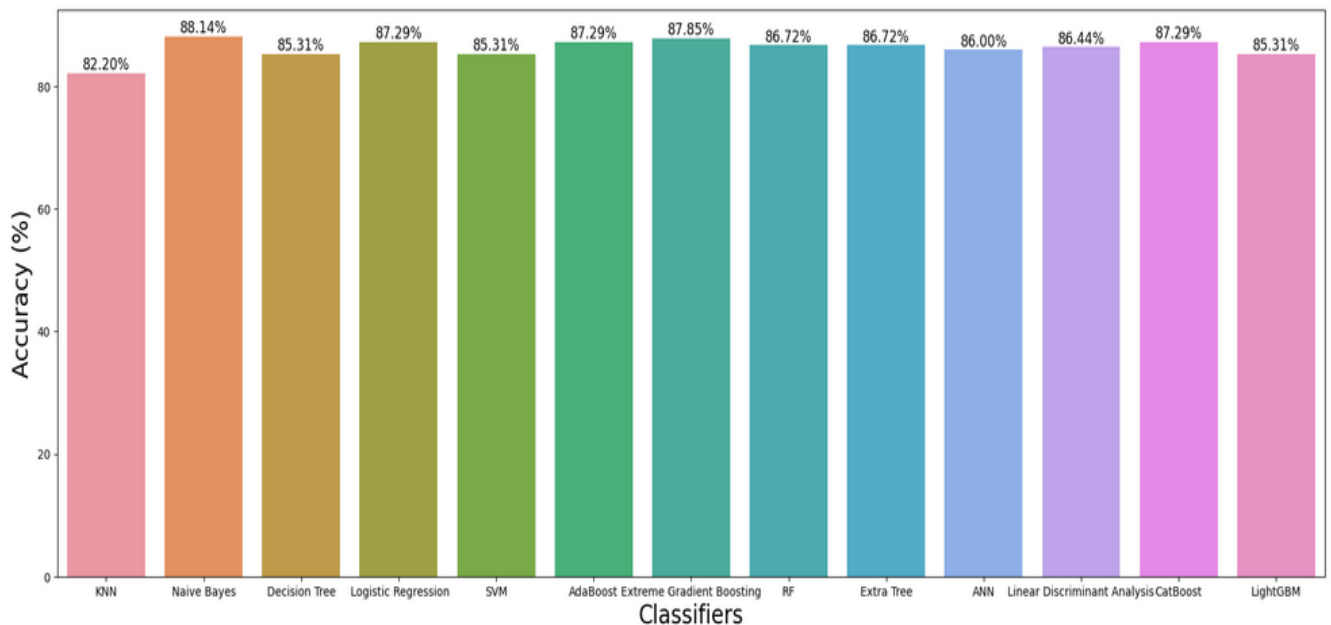


Figure 44 The Distribution Map of the Models Accuracies for Dataset1

From the above graph, which includes all accuracies of the models applied in this project for the first dataset, it is clear that the highest score of accuracy was given by the Naive Bayes model and the accuracy recorded was 88.14%. The 88.14% is the

highest score, which indicated that the best machine learning model performed in this project and for this specific dataset was Naive Bayes. However, the other dataset showed different results, which shows that the prediction results of the models always depend on the dataset.

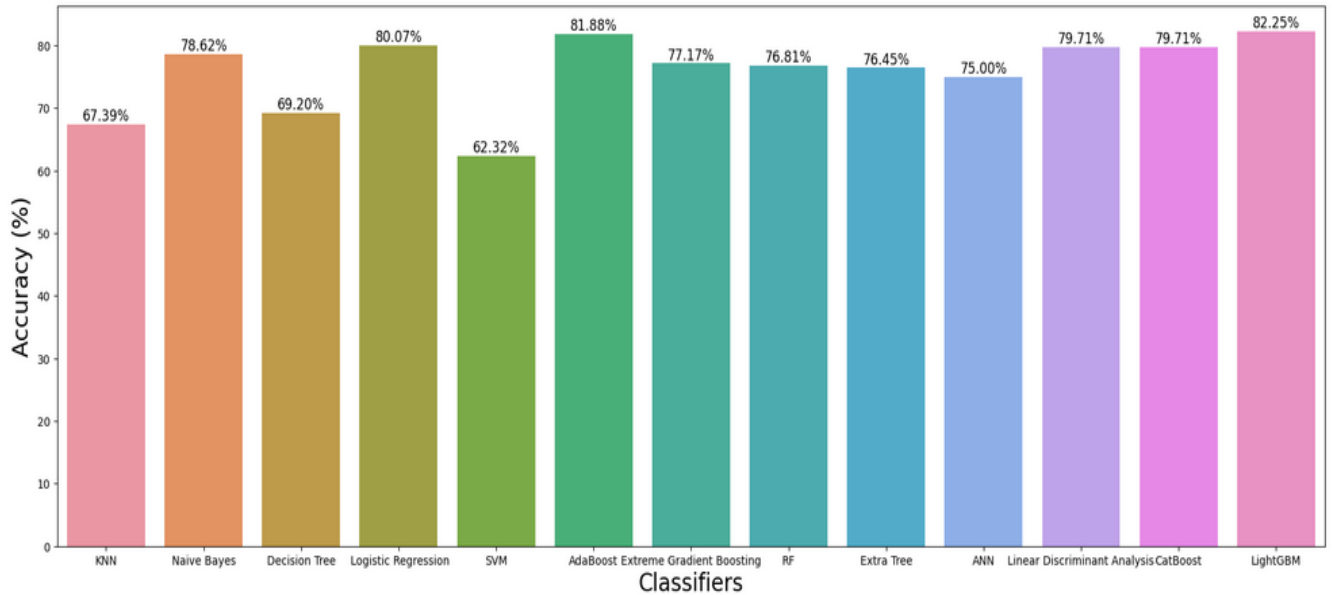


Figure 45 The Distribution Map of the Models Accuracies for Dataset2

According to the figure above, which includes the prediction results of the machine learning models for the second dataset. It is clear that this dataset not only showed lower percentages of accuracies compared to the first dataset but also the machine learning model that showed the highest accuracy for this dataset was the LightGBM model, with an accuracy of 82.25%. For the first dataset, the lowest accuracy was recorded from the K Nearest Neighbor, which is 82.20%, and it is a close result to the highest accuracy recorded in the second dataset. For this, it is obvious that the first dataset is better when applied to the machine learning models than the second dataset in the prediction of heart failure. In addition, since the accuracies were different and the models that gave the highest accuracies in the two datasets were also different, other methods were applied to get more details on how these models worked, and one of these methods was a confusion matrix. The confusion matrix was applied to the two models, which gave the highest accuracies in the two dataset, and these models were also compared according to their performance.

In addition, after the models were built and their accuracies were collected and compared, an extra method was applied to get a clear idea about the training and tested sets, and whether they worked well on the machine learning models or not. The method was the confusion matrix, which is also known as the error matrix. As mentioned by , the confusion matrix, as known by its name, is a matrix that can give a summary of the machine learning models' work when it is working on a test date set, and this was the use of it in this project. The confusion matrix has four matrices presenting the work of the models. The first part of the matrix, which is called True Positive, represents the times that the machine learning models predict correctly, which showed in this project how many times the machine learning models predicted sick people with heart failure correctly. The second part, which is called False Positive, represents the times that machine learning models predicted healthy people as sick people with heart failure. The third part, which is called False Negative, presents the times that machine learning models predicted sick people with heart failure who were not sick. Lastly, the fourth part, which is called True Negative, presents the times that machine learning models

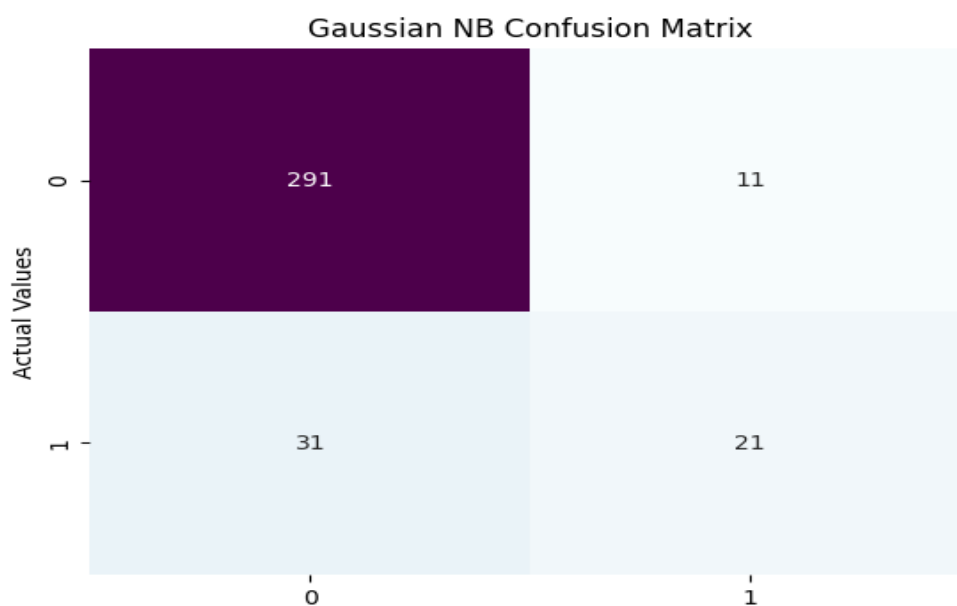


Figure 46 The Confusion Matrix of the Naive Bayes Model for Dataset1

predicted healthy people as not sick people. In addition, the confusion matrix also gives a report that explains how the models worked and a detailed explanation of the work of the confusion matrix.

According to the above figure, which indicates the confusion matrix for the Naïve Bayes model from the first dataset. The matrix shows clearly that the Naive Bayes showed good performance in predicting the people with heart failure correctly, and there were 291 instances that the machine learning model was correct in predicting the people with heart failure. On the other hand, the results from the false positive showed that the model had some wrongly predicted patients, which showed 11 instances that the machine learning model predicted some healthy people as the heart failure patients, and this could indicate that the model might had some noise or it could be over sensitive. Additionally, the matrix showed that the model gave 31 instances as False Negative, which means that the model showed that the patients with heart failure were not predicted as patients in 31 instances, and that could have happened because the model might have missed some important patterns while predicting. Lastly, the matrix showed that the model got True Negative in 11 instances, which means that the model did not give good performance in the True Negative, where it only predicted healthy people as people with no heart failure in 11 instances. The result recorded from the True Negative showed that the model is not performing well when it comes to identify the negative instances.

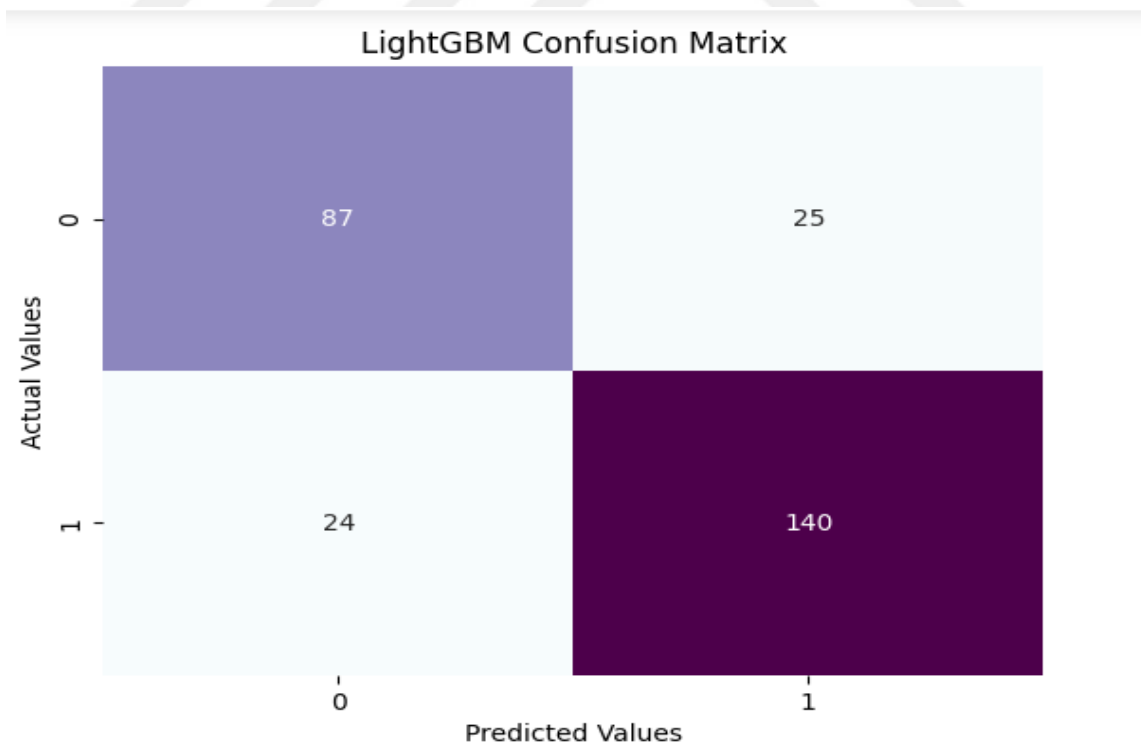


Figure 47 The Confusion Matrix of the LightGBM Model for Dataset2

In comparison to the confusion matrix of the Naive Bayes model from the first dataset, the figure above shows the confusion matrix of the LightGBM model for the second dataset, which gave the highest accuracy performance according to the other accuracies from the other models for dataset 2. The confusion matrix in the figure shows how the LightGBM model performed while predicting, and it is clear that the performance of the model was not good enough. According to the confusion matrix, the model showed 87 instances for the True Positive, which means that the model predicted the patients with heart failure correctly in 87 instances but could be better since it is important to the model to score the correct predictions in more instances. On the other hand, the results recorded for False Positive showed that the model had some wrongly predicted patients, as the False Positive was 25, which indicates that the machine learning model predicted some healthy people as patients with heart failure in 25 instances and that could be a result of noise or that the model is over sensitive.

In addition, the matrix also showed that the wrong prediction with the False Negative, which means that the machine learning model showed a prediction of some patients with heart failure as healthy people in 24 instances, and this could be happened because the model might have missed some patterns or features while predicting. The last part here of the confusion matrix for the LightGBM model is the True Negative that scored 140 instances, which indicates that the machine learning model predicted healthy people as people with no heart failure in 140 instances, and this indicates that the model gave good performance when predicting healthy people. However, the confusion matrix did not give a good report for the performance of the LightGBM model since it only gave good performance when predicting True Negative, and this might be affected by the dataset or the noise that could happen when training and testing the model.

As a result of the confusion matrices above, the prediction of the models in both datasets was applied to some methods to enhance their prediction and to improve the performance of the models. The first method applied to the models was the cross-validation method, which is known as a method that helps in improving the models' performance. Cross-validation was applied to all the models in both datasets in the same way, and the following steps were applied to all the models for both datasets.

Before starting the cross-validation method, it was essential to ensure that the results from the cross-validation method were collected in a specific list, and to do that, a list was made to contain all the cross-validation results from all models used in this project. To start with the cross-validation method, the list was made and was given a name `crosslist`, and this list will be the returning list for all the models used in both datasets. Moreover, the cross-validation method was imported from the library `sklearn.model_selection`, which also included `KFold` and `StratifiedKFold` that they were also imported in the notebook. It is crucial to mention that to be able to use cross-validation, the `k` folder needs to be imported since the main job of the cross-validation is to divide the dataset into these folders.

Moreover, the folders were assigned to be in the cross-validation, which would allow the method to give a better and more reliable estimation of the models' performance. It is known that the first step in using the cross-validation method is to put the dataset into these folders equally then the models were applied to be trained and validate 8 times, according to the folder number, and each time the model was used one folder to validate and the other 7 folders are used to train the model. In addition, after training the model and validating it, the next step in cross-validation is to evaluate the performance of the model, which is calculated according to the predictions from the previous set, which is the training and validation step. Lastly, the final step in the cross-validation method is to get an average for the performance of the models, which is by getting the average of all performances of the models from the previous step.

Furthermore, in order to apply the cross-validation method to the models, it was crucial to consider starting to build the method by including the wanted model, then including the features in the `x` variable and the target from the `y` variable, including the number of folders, which is 8 in this case, and lastly include the scoring, which was chosen as accuracy for this project since the project is about finding the best accuracy of prediction. Additionally, all the steps mentioned above were applied to all the models that were used in the project, and all the accuracies, which were contained from the models after they were applied to the cross-validation method, were included in the `crosslist` to compare their results.

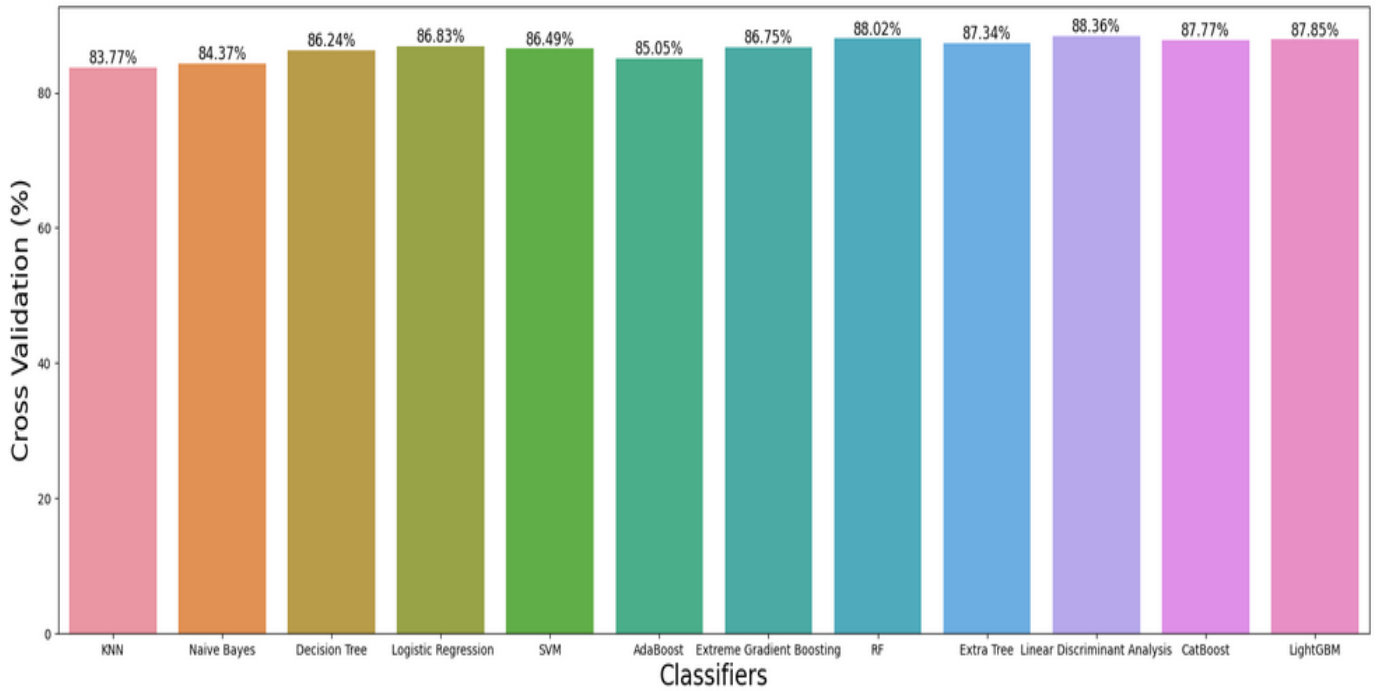


Figure 48 The Distribution Map of the Cross-validation Accuracies for Dataset1

The figure above shows the percentage of accuracies of the models for the first dataset after they were applied to the cross-validation method. The results were quite good since the cross-validation method had the ability to enhance some of the machine learning models; as an example, it is clear that the cross-validation method improved the predication percentage of the Linear Discriminant Analysis model from 86.44% to 88.36%, which is quite a good improvement. On the other hand, the cross-validation method did not show improvement in the Naive Bayes model, which gave the highest accuracy before it was applied to the cross-validation method. The Naive Bayes model was affected negatively by the cross-validation method, in which its accuracy was dropped from 88.14% to 84.37%, which is considered extremely low accuracy. The change in results might happen due to the over-fitting or under-fitting or the randomness while splitting the dataset when using the splitting method. However, using the cross-validation method ensures that issues like the over-fitting and the under-fitting do not happen, and having the average from the folders in the cross-validation method gives a more reliable estimate of the model's performance.

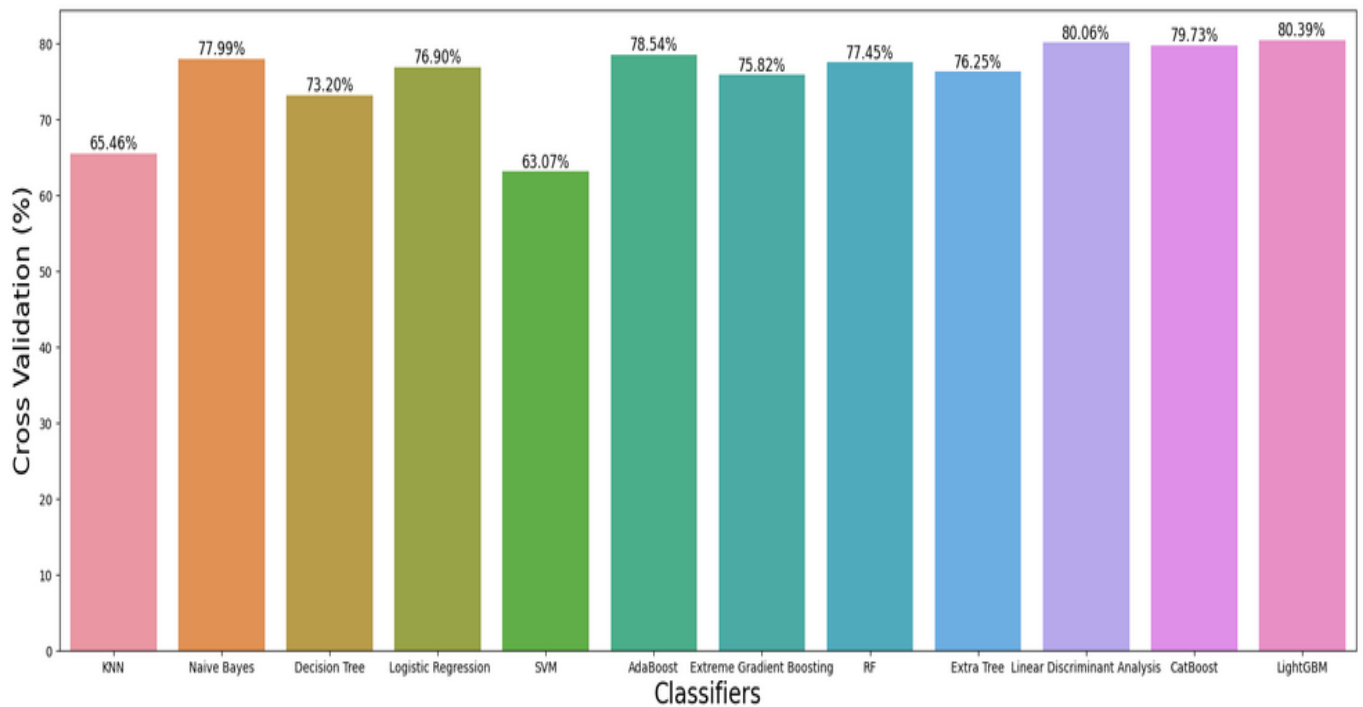


Figure 49 The Distribution Map of the Cross-validation Accuracies for Dataset2

The figure above shows the percentage of accuracies of the machine learning models from dataset 2 used in this project after they were applied to the cross-validation method. The results showed a slight change in the accuracies of the models, and it did not give much improvement in some models' accuracies; as an example, it did show some slight improvement in the Linear Discriminant Analysis model since it enhanced the model's accuracy from 79.71% to 80.06%, which means that the Linear Discriminant Analysis had some noise or over-fitting when the model was calculating the accuracy. On the other hand, most of the models showed some drop in their accuracies, like the LightGBM model, which gave the highest accuracy when compared to the other models, but when applied to the cross-validation method, it showed some significant drop, and that was from 82.25% to 80.39%; however, it is still giving the highest accuracy when compared to the other models after the cross-validation method was applied. Normally, the cross-validation method enhances the machine learning models' accuracies more than using the simple splitting method, and that is because of the way it works, in which it splits the dataset into a number of folders equally for training and testing, and this results in giving a strong evaluation of the machine learning models' abilities.

Furthermore, after getting the results from the cross-validation method, it was vital to try another method to enhance the machine learning models' accuracies in order to know what was the problem with the cross-validation results that made it not to show much improvement to the accuracies. At this point, the Principal Component Analysis was applied to all machine learning models to get a more detailed idea about the problems that prevent the enhancements for the machine learning models' accuracies. The Principal Component Analysis was used with both datasets' machine learning models in the same way, and the following steps were applied to all machine learning models in both datasets.

Before starting the Principal Component Analysis method, it was essential to ensure that the results from the Principal Component Analysis method were collected in a specific list, and to do that, a list was made to contain all the Principal Component Analysis results from all models used in this project. To start building the Principal Component Analysis, a list was made and given a name, pcalist, which was the return list of all the accuracies that were collected from the machine learning models.

Furthermore, to be able to work with Principal Component Analysis, the number of components was specified in order to only keep a specific number of the principal component from all the times that the principal component was generated. The chosen number of the component was to help to balance between reducing in the dimension of the data while containing a sufficient amount of information to represent the total dataset, and according to that, the component number was assigned to be ten components, which ensured that there was not any loss of the essential information as well as not faced over-fitting. After the number of components was assigned, the principal component analysis was built by starting to import the StandardScaler technique, which was assigned to the variable scaler. After that, the scaler variable was fitted to the training set that is included in the x-train variable, which contained features, and then transformed by the transform method, which is considered a method that calculates the mean and the standard deviation of the training data in the x-train variable.

In addition, the scaler variable was also fitted to the testing set that is included in the x-test variable, which had the remaining features and then was transformed by the transform method, which then calculated the mean and standard deviation of the testing data in the x-test variable. Moreover, the next step was to create a variable with

the pca name, which included the PCA object together with the number of components that are used to contain a reasonable amount of variance while transforming the data into lower dimension space. To be able to use the training and the testing data with the pca method, the pca was fitted to both training and testing variables, which were applied to the scaler method previously, and fitting them to the PCA was to make the PCA method learn from the eigenvalues and vectors, which are included in the training data.

In addition, the same fitting method was used for the testing data from the scaler method, which was used to be fit to the PCA method to help the method learn from the features in the training set of the scaler. It was important at this point to ensure that both training and testing were transformed in the same way and by using the same methods because this would help with the consistency of processing. Lastly, after finishing the preparation of building and fitting the Principal Component Analysis, it was applied to all the machine-learning models used in both datasets, and the accuracies of these models were collected in the assigned list, which is pcalist.

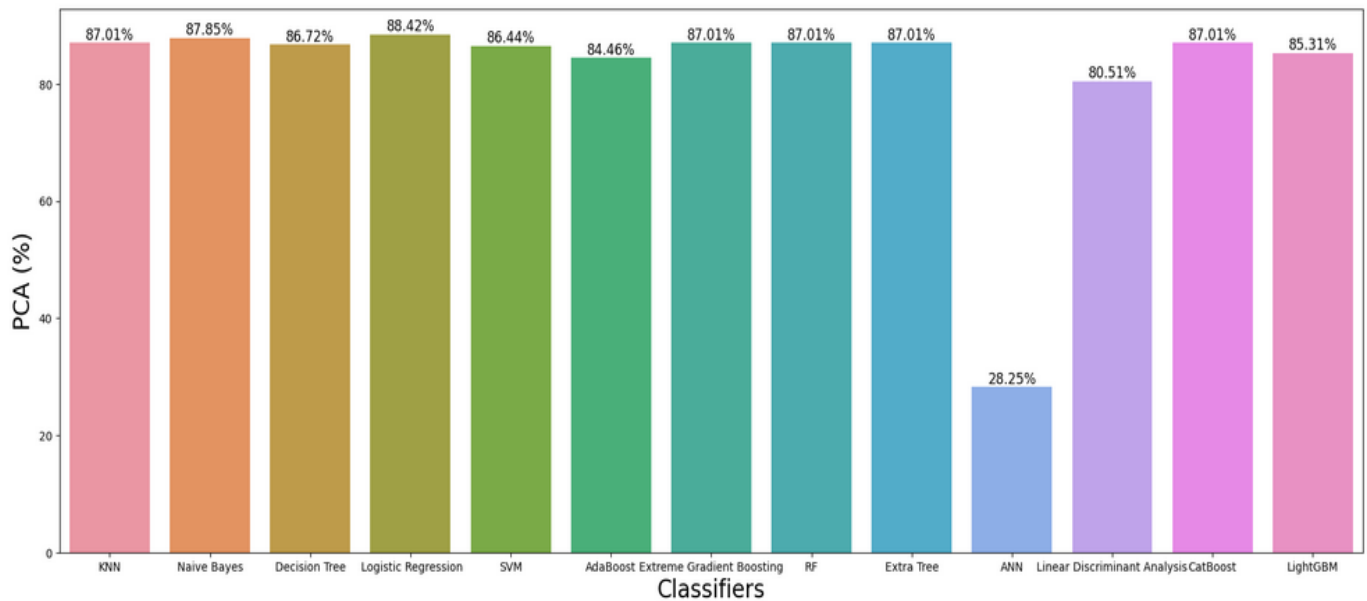


Figure 50 The Distribution Map of the Principal Component Analysis Accuracies for Dataset1

As can be noticed from the figure above, which shows the percentage of the machine learning models' accuracies for dataset 1 after they have been applied to the Principal Component Analysis that there was some slight increase in the percentage of their accuracies when compared to the results from the cross-validation method. In general, the highest accuracy gained from applying the principal component analysis method was recorded from the Logistic regression model, which was 88.42%. As a result, it is considered the highest accuracy that the Logistic Regression model gave in comparison to the normal accuracy and the accuracy with the cross-validation method. The logistic regression model scored 88.42% when applied to the pca method, and this is considered the highest accuracy scored in this project. However, when applied to the cross-validation method, it did not show good accuracy, which was 86.24%, while when the model gave its accuracy without any methods included, it did give something in between the pca method and cross-validation method, which was 87.29%. On the other hand, the Naive Bayes model showed at the beginning the highest score of accuracy when it was working without adding any methods to it, which was 88.14%, but when it was applied to the cross-validation method, the accuracy of the model decreased significantly and reached the percentage of 84.37%, which as one of the lowest accuracies scored from the cross-validation method; however, the model had an increase in the accuracy when it was applied to the pca method, which reached the score of 87.85%.

In addition, the Linear Discriminant Analysis method scored a poor number when it was applied to the pca method, which was 80.51%, as can be seen from the figure; however, it showed the best and the highest accuracy when applied to the cross-validation method, which was recorded as 88.36% and it is considered the highest accuracy that the model scored among the three methods because when applied to the splitting method it only gave 86.44% of accuracy. The changes in the accuracies that were recorded from the models when applied to the method are considered normal since the methods work in different ways, and the differences between the methods are the part that makes the models work effectively. Lastly, it is important to mention the lowest accuracy recorded from the models when they were applied to the pca method, which was given from the Artificial Neural Networks that gave 28.25% of accuracy and that is considered the worst accuracy. However, the cause of the low accuracy from ann model might be from some loss of information since the pca method work

by lowering the dimensional spaces, which might cause loss of information and that would definitely affect the work of the artificial neural networks model, which needs a vast data to be able to learn complex patterns effectively.

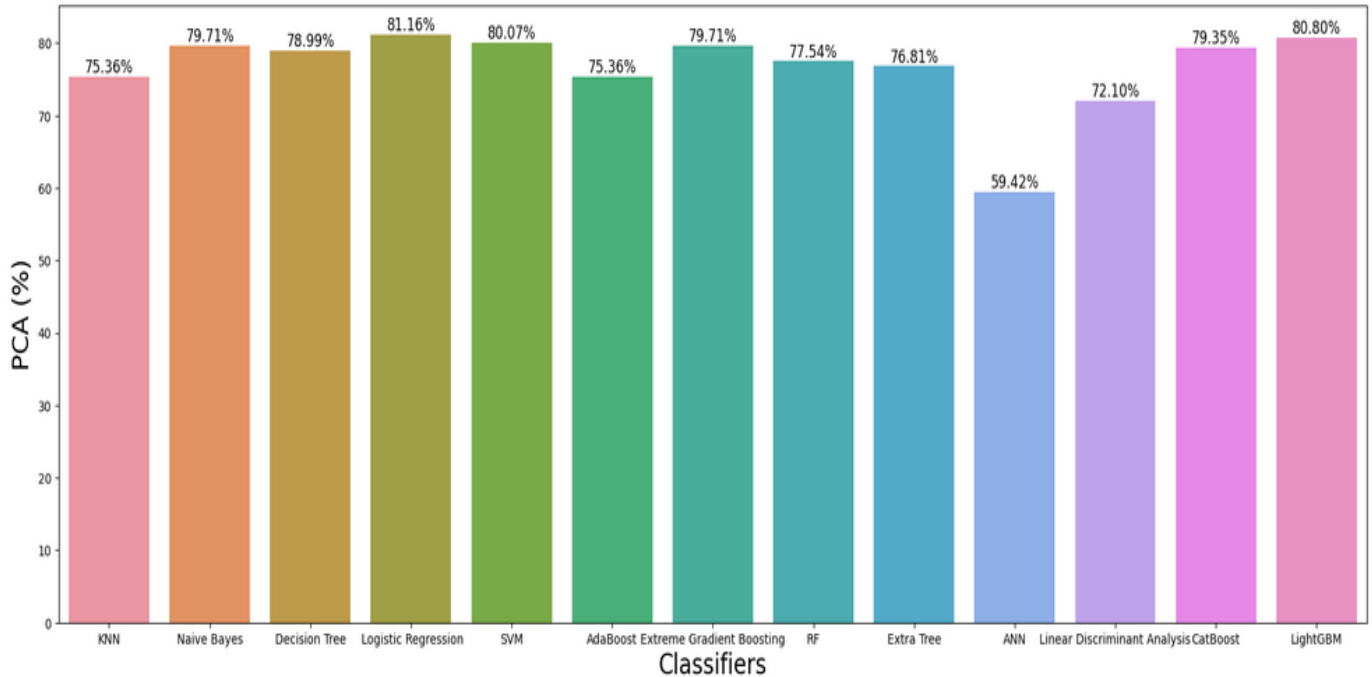


Figure 51 The Distribution Map of the Principal Component Analysis Accuracies for Dataset2

The figure above shows the accuracies recorded from all models in dataset 2 after they were applied to the Principal Component Analysis method, which showed some increase in the accuracies for almost all the models when it is compared to their accuracies while using the cross-validation method or the splitting method. It is clear that the highest accuracy given among the models was from the Logistic Regression, which is similar to the first dataset when the model also gave the highest accuracy after it was applied to the Principal Component Analysis method; although the model did not show a useful accuracy percentage when it was applied to the cross-validation method in the dataset 2, which was 76.90%. In fact, the Logistic Regression model showed a significant drop when it was applied to the cross-validation method in comparison with the accuracies that were given by the model when it was applied to different methods, as it gave 80.07% of accuracy when it used the splitting method, and that indicated that the Logistic Regression model worked better when applied to the pca method.

In addition, the LightGBM model, which gave the highest accuracy when recording its accuracy after using the splitting method, which was 82.25%, showed a slight drop when applied to the pca, since it gave a record of 80.80% of accuracy; however, the accuracy result that was recorded from the LightGBM while it was applied to the pca was a bit higher than the accuracy recorded for the model when applied to the cross-validation method, which was 80.39% accuracy, and this means that the LightGBM worked better when it used the splitting method. Lastly, it is essential to mention that the ann showed a considerable drop when applied to the pca method as it did in the first dataset, which recorded an accuracy of 59.42%, and that is considered as an unreliable accuracy, which shows that the Artificial Neural Networks method does not work well when it is applied to the Principal Component Analysis.

3.3.4 Reliability and Validity

In this section, the reliability and the validity are discussed since they were essential in developing the machine learning results of predicting heart failure in the two datasets included in this project. It was important to consider applying some examinations on the reliability and validity of the model predictions that were used in this project to know the way these models work, whether they can keep giving acceptable accuracies after the examinations, and how to improve their accuracies to gain better prediction for heart failure. At first, to be able to help in applying the reliability to the prediction of heart failure, some methods were used on the models in this project. The first method was used for getting the prediction from the machine learning methods was the splitting method. The splitting method is a method that splits the dataset into two subsets, which are the training subset that is used to train the models and the testing subset that is used to test the models after it is trained. Moreover, this method only splits the dataset one time, and after that, it gives one accuracy for each model as a result of the way it was trained from the dataset. This method was applied to all the models included in the dataset, and the results were obtained from the work of this method showed that the models' performance was stable for both datasets.

In addition, the second method applied to obtain the reliability is the cross-validation method, which is known as a method that is used to help in improving the machine learning models performance, and, as a result of that, it helps get the models

accurate prediction. The cross-validation method usually separates the dataset into k-fold parts, which were assigned to be ten folders for this project, and these folders contain equal parts of the dataset. Then, it uses these parts in training and testing the model by assigning one folder each time to be the testing folder, and the other nine folders are assigned to be the training folders, and this process is repeated for each folder. After that, the performance of the training and testing was evaluated, and the average of the ten outcomes was taken to be the reliable accuracy from the cross-validation method to all models. The cross-validation method gave more reliable results than the splitting method for both datasets, but the first dataset is the dataset that gave higher accuracies for predicting heart failure, which made it more reliable than the second dataset in predicting heart failure.

Lastly, the last method used to assess the reliability of the heart failure prediction models is the Principal Component Analysis, which was also used to help improving the machine learning models' accuracies. The principal component analysis method mainly works by reducing the data to a low-dimensional space while keeping the crucial information. The method also uses a specific number of components, which is assigned in this project as ten, and it indicates the number of times that the principal component will reduce the dimension of the data to ten dimensions. The Principal Component Analysis showed not only a slight increase in the model accuracies but also a significant increase in the accuracy of the Linear Regression model in both datasets, which showed that it is a reliable method and worked well with the Linear Regression model; however, it showed a massive drop in the Artificial Neural Networks model for both datasets, which indicated that the pca method does not work well with the ann model. In general, the accuracies gained from the method applied to the machine learning models were reliable, and the most reliable accuracy gained from the models was 88.36%, which was given by the model Linear Discriminant Analysis after it was applied to the cross-validation method.

Additionally, to be able to check the validity of the models used for prediction, some approaches were considered to be used in this project. The first approach that was considered was checking the features involved in both datasets and how they are related to heart failure. The features were studied one by one to ensure that they have the factors that could lead to heart failure or affect the heart negatively, and were also

compared to the datasets included in the literature to determine whether they had similar features and if these features are related to heart failure. It is essential to mention that the studies included in the literature were mostly about heart disease since there were few studies applied on heart failure when the research was done; however, it is important to state that most of the factors that lead to heart failure or affect the heart negatively are similar to the factors that lead to the heart diseases since the heart failure is one of the four kinds of heart diseases.

Lastly, the other approach that was used to check the validity of the model is using the area under the curve to analyze the work of the model after getting its accuracy and to see if the accuracy with the used features is acceptable when compared to the accuracies from literature or not. The area under the curve usually shows the most details for how the machine learning models perform, so it was important to use the area under the curve to have a detailed idea about the performance of the model.

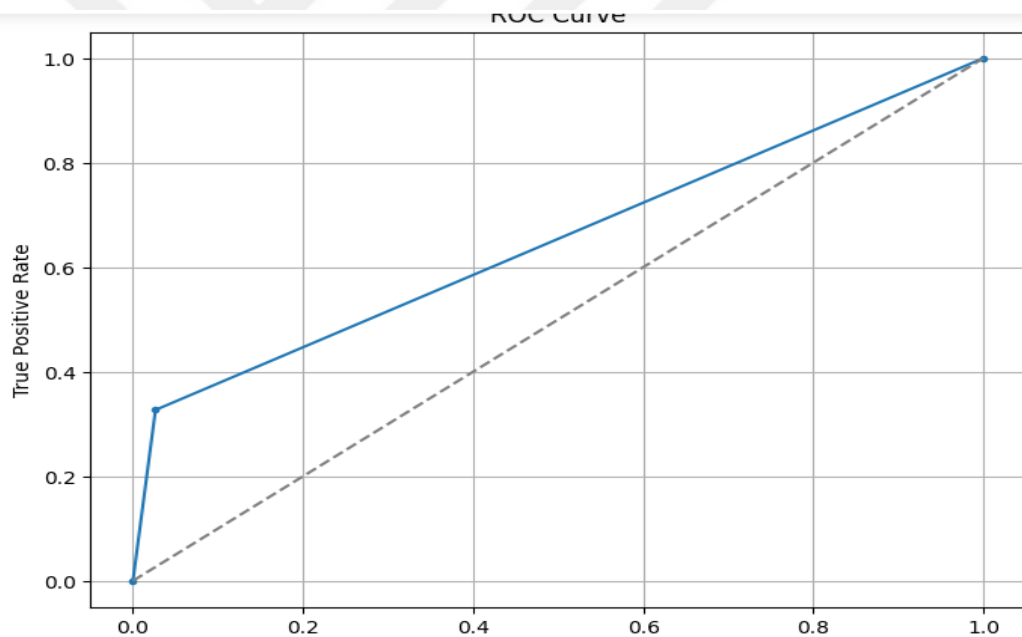


Figure 52 The AUC-ROC of Linear Discriminant Analysis model for Dataset1

The figure above shows the area under the curve of the Linear Discriminant Analysis model, which gave a reading of 0.6502165053489557. The results from the area under the curve were quite good since 0.6 in the area under the curve is considered

to be a good result, which means that the model is working in a good way but still needs to enhance its performance, and that would give a more accurate prediction.

In general, the accuracies given from the machine learning models used in this project from the three methods that were applied to the models to get the best accuracy gave quite good results for the first dataset, when compared with the results given in the literature. The highest accuracy given in the literature was 96%, as mentioned in (Krishnani, Kumari, Dewangan, Singh, & Naik, 2019), which is a reliable result for prediction. In addition, the highest accuracy from the machine learning models used in this project for dataset 1 was 88.36%, which was given from the Linear Discriminant Analysis model after applied to the cross-validation method, which is also reliable result according to the dataset 1 used for this project. It is essential to mention that the dataset from the literature was made to study and predict heart diseases, and the two used in this project were to study and predict heart failure. This might be the reason behind the differences in the accuracies since the two projects are applied to different datasets. As a result, it is possible to say that the accuracies of prediction for dataset 1, which were collected from the machine learning models after they were applied to different models, are acceptable and are reliable according to the given dataset.

3.4 Limitations

This study has provided reliable results in terms of using machine learning algorithms in predicting heart failure; however, there were some issues and limitations that were faced during the process of prediction. First of all, the first issue was faced is the limitations in the quality and lengths of the datasets. As it is known in machine learning that the models totally rely on the characteristics of the features given in the dataset and their relation with the required target, which is predicting heart failure in this case. Although both datasets studied in this project included features that are related to heart failure, the records in both datasets did not have enough quality to give the machine learning models the ability to give more accurate predictions. For the first dataset, there were some missing values from most of the features, and they were filled with the mean of these features, which is a typical method to fill in missing values when applied to machine learning models. It is important to mention that when the first dataset was collected, it was applied to an analysis that ensured that the features in the dataset did not include more than 25% of missing values, as mentioned in (Zhou, et al., 2021). However, even after filling in these values, it did have an effect on the

function of the machine learning models, which might be caused a decrease in the accuracy predictions.

For the second dataset, there were no missing values from the features in this dataset; however, there were some limitations in the length of the dataset, as there were only 918 records of patients' information, which is considered to be a bit low for studying and predicting heart failure. In addition, there were some limitations in the features for the second dataset, as it only included 12 features, which might had affected the training and testing part in the machine learning models; on the other hand, the first dataset included more features that were related to heart failure, which was 49 features and this might be the cause of the differences in the performance in machine learning models as well as the length of the first dataset that included a record of 1177 of patients' information, and this also could have affected the results and could be the reason behind the higher accuracy of predictions that were given from the first dataset when compared to the accuracy of prediction that were given from the second dataset. It is crucial to state that the length and the features of the dataset do affect the function of the machine learning models and it affects their accuracies, and for this project, it was shown from the comparison between both datasets' accuracies the effect of the length and the quality of features.

The other limitation considered is some loss of information or over-fitting that happened when the method of machine learning was used for prediction. During the process of building the machine learning models to predict heart failure, they were applied to some machine learning methods to be able to get the prediction accuracies. The first method used was the splitting method, which is considered a simple method to use for splitting data into two sets, and these subsets are used in training and testing the machine learning models. The splitting method is useful when providing a fast prediction from the machine learning models' performance; however, using this method could cause over-fitting or randomness in splitting, which might have happened when this method was used during this study when it was applied to all machine-learning models used for both datasets.

The second method used in this study is the cross-validation method, which was applied after the splitting method was applied in order to enhance the machine learning models' prediction accuracy and their performance since this method relies on the idea of dividing the dataset into a specific number of times. The cross-validation

works by dividing the dataset into a number of folders, which was ten in this case, and that number of folders works by making one of these nine folders a test data and the other nine as training data, and this process repeated to all the ten folders. The way the cross-validation method works gives a more reliable prediction of machine learning models than the prediction from using the splitting method since it uses all the datasets in testing and training the models, which was the main reason for using the cross-validation method. The cross-validation method gave better predictions from machine learning models for both datasets; however, a couple of the models showed better performance with other methods, which might be from the complexity of these models that the cross-validation method did not work well with them.

The last method was used is the Principal Component Analysis method, which was also applied to all machine learning models in both datasets, and that was to get better predictions from the models. As this method works by minimizing the dimension of the data from high to low-dimensional data, which could be useful in enhancing the effectiveness of the models' performance and help in the identification of the patterns for the high-dimensional data. The use of principal component analysis did enhance the performance of some of the models used for both datasets; however, it made some significant drop in the performance of the other models, which could be a result of the loss of some information while the method minimizes the dimensional of the data or it could be by assuming the linear relation between the features, which might not be correct for some of the features.

In general, the use of the mentioned models and the type of datasets chosen for this project gave reasonable and acceptable results, but there were some limitations that were impossible to avoid when working on the project. It is essential to consider the use of the mentioned methods because they are an excellent method to help the machine learning models predict and give reliable accuracies; however, there will always be a part of the disadvantages to consider while using the methods. In addition, the limitations faced in this project did affect the machine learning models' accuracies but did not prevent the models from giving reliable accuracies when compared to the other studies from the literature.

Chapter 4

Findings

In this section, the findings of this project are discussed by explaining what are the results of using machine learning models to predict heart failure. There are lots of things that were found during this study. The study focused on finding the best machine learning model for predicting heart failure on two datasets, and these two datasets were included carefully since the study is about comparing the two datasets' results and seeing if the length and the validity of features would affect the results of prediction. The first dataset was chosen to be the larger dataset, which included 1177 patients' information records and had 49 features that are related to heart failure. Moreover, the features of the dataset were analyzed to get a general idea about them, and it was clear that the average of the patients' ages was 74 years old, which was acceptable since, mainly, the risk of having heart failure increases when the person gets older. The study also included data analysis and visualization for all the features that were found to be correlated to the outcome after the dataset was applied to the heat map, which gave a confusion matrix of the features and showed which features are the most correlated to heart failure. It was found that the age, atrial-fibrillation, heart-rate, respiratory-rate, MCH, MCV, and many other features from the first dataset are related to heart failure. According to that, all the features were analyzed and visualized versus heart failure to get a clear idea of how these features are related to heart failure.

In addition, the dataset was applied to the machine learning models to get the models' accuracies of prediction and to get more effective accuracy; the dataset was applied to three kinds of methods to get the best accuracy for predicting heart failure. It was found that the highest accuracy from all the models, and after these models were applied to the different methods, recorded from the Linear Discriminant Analysis, which gave a prediction accuracy of 88.36%, and that is after the model was applied to the cross-validation method. It is essential to mention that there was a model that gave higher accuracy when it was applied to the Principal Component Analysis, which is Logistic Regression, with an accuracy of 88.42%; however, the model gave 86.83% when applied to the cross-validation method, which indicates that the Logistic Regression model worked well when applied to the Principal Component Analysis method.

On the other hand, the second dataset was chosen to be the smaller dataset, which had records of 918 patients' information with only 12 features that are related to heart failure. The features of this dataset were also analyzed to get an idea about them, and it was found that the average age of patients in this dataset was 53.5 years old, which was acceptable as well since the average still included in the age group of elders. The dataset also was applied to data analysis and visualization for all the features it has. Also, it was applied to the heat map, which gave a matrix of the features correlation, and from the map, it was possible to indicate which features are more correlated to heart failure. It was found that the age, gender, resting blood-pressure, fasting blood-sugar, and other features were related to heart failure. As a result of that, these features were studied, analyzed, and visualized together with heart failure to gain a clear idea of how these features are related to the heart failure. Additionally, this dataset was applied as well to the machine learning models to get the models' accuracies of prediction, and this dataset was also applied to the same methods to enhance the accuracy of the machine learning models.

It was found that the highest accuracy was recorded from all the models used in this project, and, after they were applied to the methods, was recorded from the LightGBM model with an accuracy of 82.25%, which was given after the model was applied to the splitting method. However, it is crucial to mention that after the models were applied to the cross-validation method and principal component analysis method to enhance the models' performance, the accuracies of the models were changed. The LightGBM accuracy was slightly decreased to be 80.39% after it was applied to the cross-validation mode, but still was the highest accuracy when compared to the other models' accuracies. On the other hand, the model that gave the highest accuracy when applied to the principal component analysis was the Logistic Regression model, with an accuracy of 81.16%, which confirms what was found from the accuracies of the first dataset that the Logistic Regression model works perfectly with the principal component analysis method.

In addition to the previous findings, some techniques that were applied to the first dataset to enhance the predictions of the machine learning models since the first dataset showed higher accuracies than the second dataset. The first technique to apply was the ensemble technique, which is used to enhance the machine learning models by combining a couple of machine learning models to get more vital prediction

accuracy. The ensemble technique used the Random Forest, the Naive Bayes, and the Support Vector Machine models together to combine them and get better prediction accuracy, and since the Naive Bayes model gave the highest accuracy among the models when applied to the splitting method, it was supposed to give better accuracy with the ensemble method. However, the accuracy given from the ensemble technique was 87.57%, which was lower than the highest prediction accuracy given in this project, which was 88.42%. The result from the ensemble technique should have been higher, but the decrease that happened to the prediction accuracy might be because of over-fitting, data quality, or even the quality of the model, which could affect the accuracy. As a result, it was found that even using some specific techniques to enhance the accuracy of prediction might not help if there was an issue with the dataset or a sudden over-fitting occurred.

The last finding that was found for this project is the kind of features that are important to the machine learning models applied to the dataset since the prediction accuracy of the models applied to the dataset did not pass the prediction accuracy of 88.42%, although the models were applied to some methods and techniques to enhance their prediction accuracy. As a result, it was essential to check the importance of the features for the models to get an idea on which of the datasets' features are most important to the models.

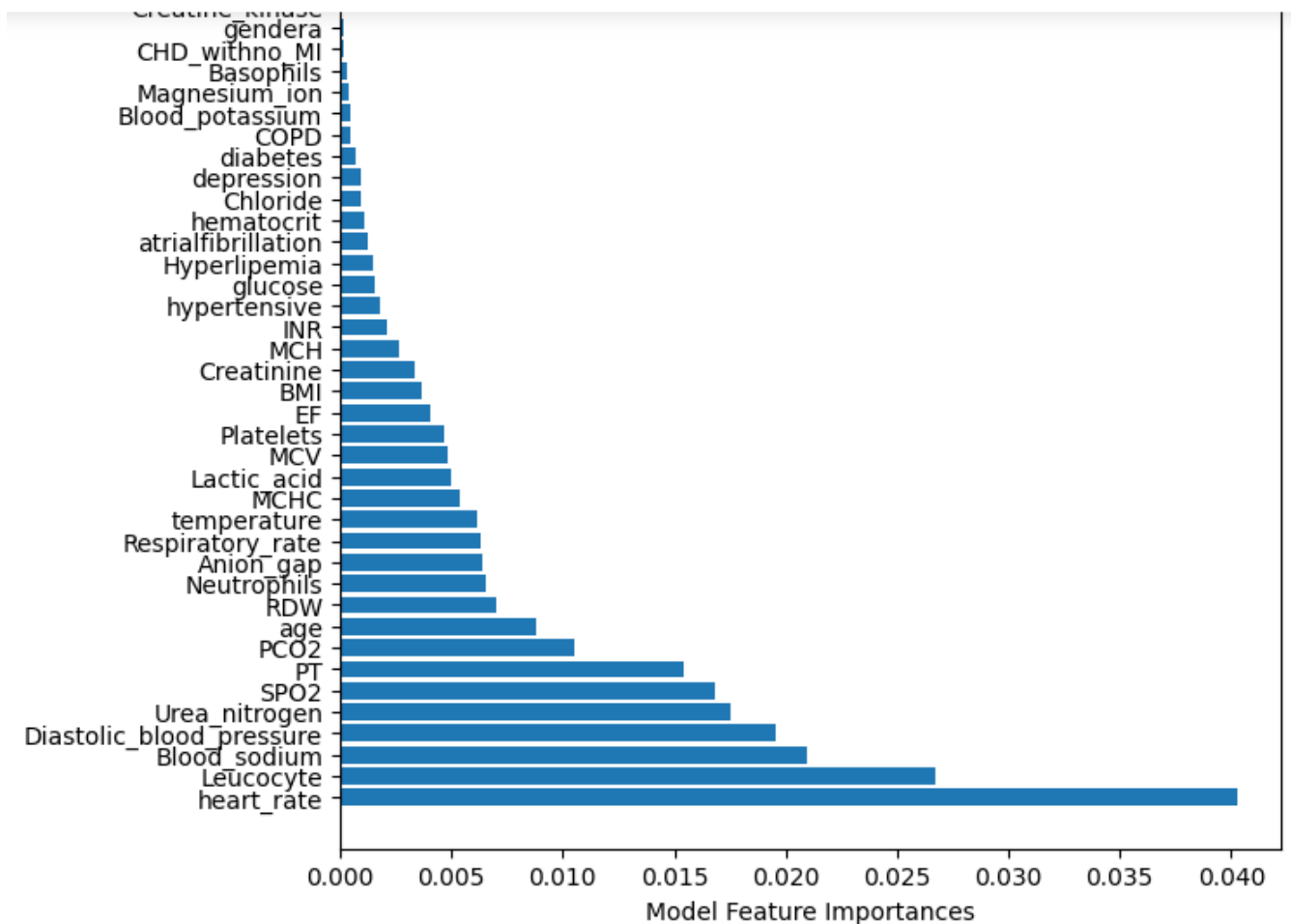


Figure 53 The Distribution map of Feature Importance for Linear Regression for Dataset1

According to the figure above, which shows the feature importance of the Linear Regression model, as it was the model that gave the highest prediction accuracy among the other models when applied to the Principal Component Analysis method. It is clear from the figure that the most important features of the Linear Regression model are the heart-rate and Luecocyte, and the least important features for the model are the CHD-withno-MI and gender. The knowledge of the important features of the model indicates that the model's most impact comes from the features that give the high scores from the feature importance distribution bar, which means that the highest feature from the graph affects the model's process the most.

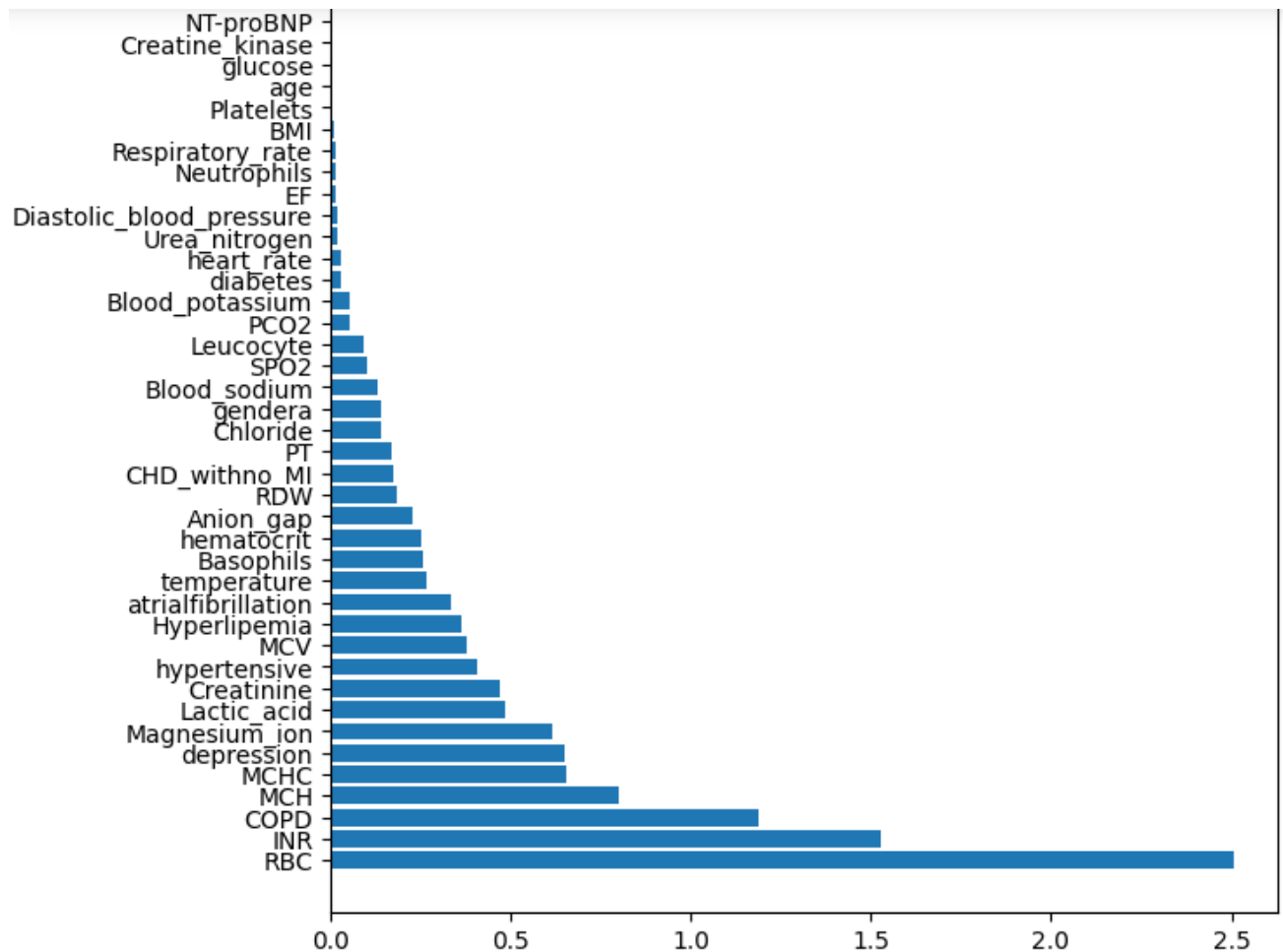


Figure 54 The Distribution map of Feature Importance for Linear Discriminant Analysis for Dataset1

The figure above shows the feature importance of the Linear Discriminant Analysis model since it was the model that gave the highest prediction accuracy among the other models when applied to the cross-validation method. It is clear that the most important feature of the Linear Discriminant Analysis model is the RBC, which is the red blood cells, which is not spotted in the Linear Regression model's feature importance graph, and that means that the feature had no effect and no importance to the Linear Regression model. In addition, the least important features were NT-proBNP, Creatine Kinase, and Glucose, which, when compared to the Linear Regression importance figure, it turned out that the Linear Regression model also was not affected by the NT-proBNP and Creatine Kinase features; however, the Glucose feature showed some slight importance for the Linear Regression model. Moreover, one of the least features that affected the model was heart-rate, although it was the most important feature that showed importance for the Linear Regression model,

which means that the models work differently on features. According to that, it was essential to compare the important features of the Linear Discriminant Analysis model and Linear Regression model with another model's important features to have a clearer idea of the features and to see if the models have almost the same feature importance or not, which will make it possible to do some feature engineering if they have similar feature importance.

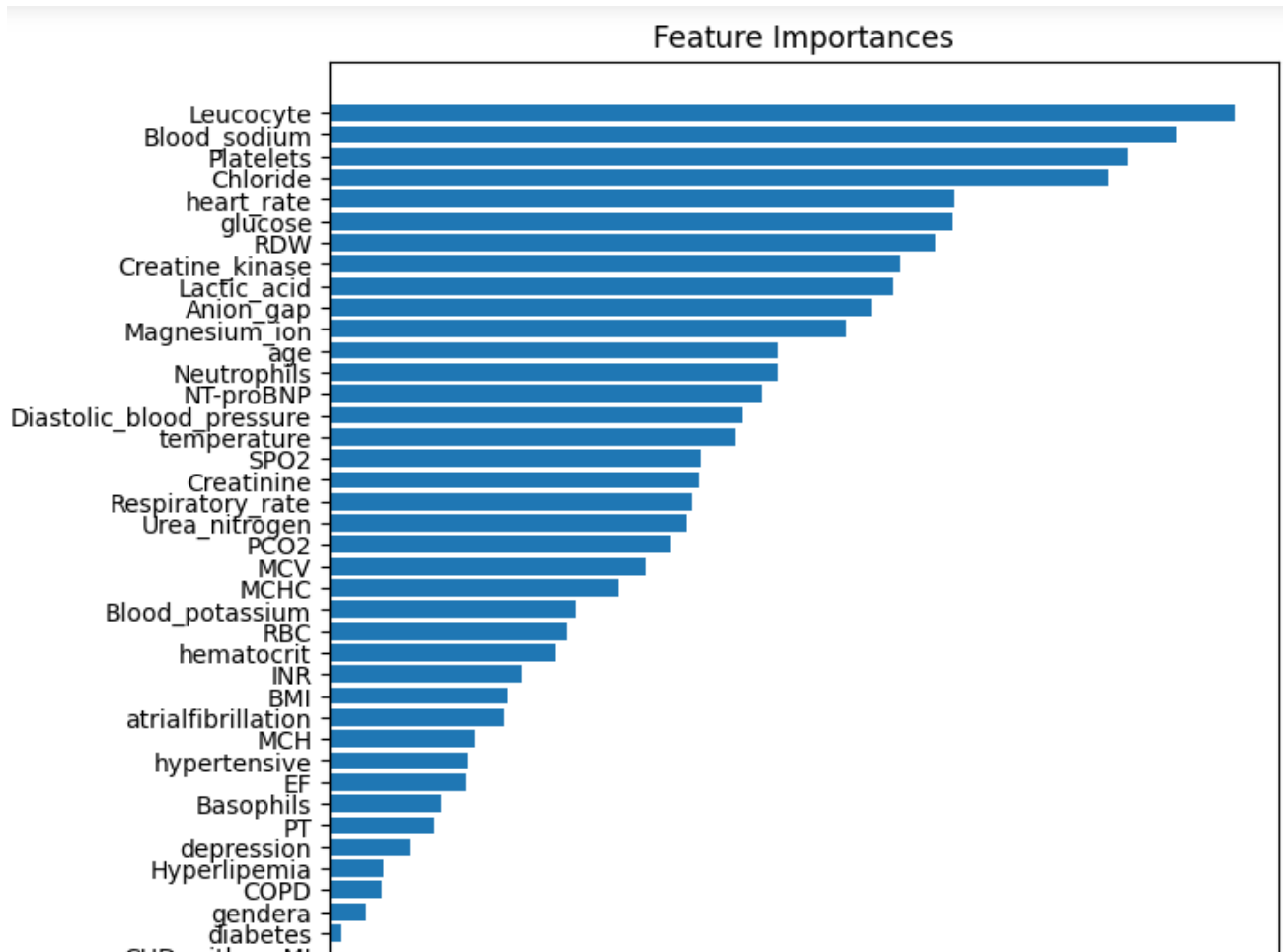


Figure 55 The Distribution map of Feature Importance for Cat Boost for Dataset1

According to the figure above that is showing the distribution map of the feature importance for the Cat Boost model that is plotted to compare it to the Linear Discriminant Analysis model and the Linear Regression model. It is clear from the map that the most important feature for the Cat Boost model was Leucocyte, which was noticed as one of the least important features for the Linear Discriminant Analysis model but was one of the most important features for the Linear Regression model. Additionally, the most important feature of the Linear Discriminant Analysis model, which is RBC, is spotted almost in the middle of the figure above, which means it is

not totally important for the Cat Boost model and not totally unimportant for the model; however, the most important feature for the Linear Regression model, which is heart-rate, was also spotted as one of the most important features for Cat Boost. In addition, the least important features for the Linear Discriminant Analysis model and the Linear Regression model are spotted at the top of the feature importance for the Cat Boost model, which means that these features are considered the most essential features for the Cat Boost model. As a result, the example model, which is the Cat Boost model, showed different importance of features from the Linear Discriminant Analysis model and some differences from the Linear Regression model, which are the models that gave the highest accuracy and that indicated that it was impossible to make feature engineering for this dataset since the models depend on different features.

In general, the findings of the study were useful in terms of predicting heart failure by using different machine learning models. Also, it was found that the use of methods that enhance the prediction accuracy for machine learning models was useful according to the machine learning models and how they work differently with these methods. Additionally, it was found that the size and the features of the dataset could affect the prediction accuracies and could either lead to high or low accuracy. On top of that, it was found that knowing the importance of the features for the machine learning models is essential if the models' prediction accuracies need to be enhanced, which will be by using feature engineering techniques.

Chapter 5

Discussion and Conclusion

This section discusses all the findings that were found during the study of predicting heart failure that was included in this project. The study was done on two datasets in which each dataset included features that were related to the heart failure, which were applied to some machine learning methods and techniques to indicate their effect on the risk of having heart failure and to obtain prediction accuracies of heart failure from the machine learning models.

5.1 Discussion of Findings for Research Questions

The study is about predicting heart failure using machine learning methods, and the findings of the study were acceptable and reliable since it was possible to predict heart failure not only for one dataset but for two datasets that were used in this study. The reason behind this study is to help medical institutions to have an idea of patients who are at risk of heart failure so that they can give them the necessary medical attention they need and the specific treatments to decrease the number of patients who might suffer from heart failure. It was essential to have an overall idea about the datasets contained in this study to ensure that these datasets would be useful to reach the target of this project, which is predicting heart failure with machine learning models. At first, some data analysis and visualization were applied to both datasets' features, and that was to understand the features and how they are related to the outcome, which was the reason for finding some missing that happened to be in the first dataset. The missing values were filled with a technique in machine learning, which uses the mean value of each feature that had missing values and fills the empty rows with the mean. After that, the data analysis and visualization showed the most correlated variables to the outcome, which were heart-rate, age, creatine kinase, leukocyte, MCV, and many other features from the first dataset. The second dataset also showed a correlation between its features and the target, which included resting blood pressure, fasting blood sugar, age, gender, and other features that do relate to heart failure medically and analytically.

In addition, both datasets were applied to the same machine learning models, which are applied according to their power of learning and predicting using the features from both datasets. The datasets were also applied to some of the machine learning methods and techniques to help enhance the accuracy of prediction for

machine learning models. The methods used were three methods, which are the splitting method, the cross-validation method, and the principal component method, and each method gave different prediction accuracy from the other method; however, all the methods showed an acceptable accuracy in predicting heart failure. After all, the machine learning models were able to predict heart failure successfully by using different models with different methods, and this was sufficient to answer the first question for this study, which was how to predict heart failure using machine learning methods.

Furthermore, it is essential to mention that the accuracies for all models using the three methods were compared to find which model is giving the best prediction accuracy for predicting heart failure. The prediction accuracies were first collected after the machine learning models were applied to the first method, which was the splitting method that works by dividing the dataset into two subsets, which presents the training and testing sets that were used on the machine learning models to train and test them. The training subset included all the features from the datasets, and the testing subset included the target from both datasets, which is the outcome. The splitting method gave different accuracies from all models for both datasets, and the higher prediction accuracy was given from the models applied to the first dataset was 88.14%, which was given by the Naive Bayes model, and the highest prediction accuracy was given from the models applied to the second dataset was 82.25%, which was given by the LightGBM model.

Additionally, the cross-validation method gave different prediction accuracies when compared to the splitting method, and this method was also applied to all models for both datasets. The highest prediction accuracy for the first dataset, which was obtained from the Linear Discriminant Analysis model with the use of the cross-validation method, was 88.36%, and the highest prediction accuracy for the second dataset that was given by the LightGBM model with the use of cross-validation was 80.39%. The last method was used to enhance the prediction accuracies of the models was the principal component analysis, which also contained different prediction accuracies for both datasets. The highest prediction accuracy for the first dataset that was given by the Linear Regression model after it was applied to the principal component analysis method was 88.42%, and the highest prediction accuracy for the second dataset that was also given by the Linear Regression model was 81.16%. As a

result, by gathering all the results obtained from the prediction procedure, it was possible to answer the second question of this study, which was “ Which machine learning method gives the best prediction accuracy?” from the results above, it was obvious that the model that gave the highest accuracy for the first dataset is the Linear Regression model after it was applied to the Principal Component Analysis model. However, the model that recorded the greatest prediction accuracy for the second dataset was the LightGBM model after it was applied to the splitting method.

In addition, it was clear from the results that included the prediction accuracies of all machine learning models that the highest accuracy recorded in this study was obtained from the first dataset, which was 88.42% prediction accuracy that was given from the Linear Regression model after it was applied to the Principal Component Analysis. The results showed that the first dataset was more suitable for predicting heart failure than the second dataset, and since this project is all about having the best-predicting accuracy for heart failure using machine learning models, it was essential to highlight that the first dataset is best dataset for this project considering that it is better than the second dataset. In addition, focusing on which dataset is better for predicting heart failure was one of the things that should be found during this study, and it was possible to reach by finding that the first dataset is better at predicting heart failure than the second dataset, and this result was the answer to the third question for this research.

Moreover, it is also essential to mention that the first dataset had more patients' records, which were 1177 of patients' information, and that is more than the patients' who were recorded in the second dataset, which were 918 of patients' information. It is possible to say that the length of the dataset could have affected the results of the machine learning models' accuracies since the length of the second dataset is less than the length of the first dataset, and from the prediction accuracy results it was obvious that all machine learning models gave higher accuracy for the first dataset when compared to the second dataset. As a result, it is reasonable to say that the length of the dataset would definitely affect the machine learning models' performance, especially since there are some models that work more effectively when it is applied to a large dataset like the Artificial Neural Networks, which was clear from the model's accuracy when it was applied to the Principal Component Analysis that it gave low

prediction accuracy because the pca model decreases the dimensional data and this leads to a loss in information.

Lastly, it was also found that the quality of the features included in the datasets did have some effect on the machine learning models' prediction accuracies, which was figured from studying the features in the first dataset and the second one. The first dataset included 49 features that are related to heart failure, while the second dataset only included 12 features that are related to heart failure. In addition, when the two datasets were applied to the heart map to check the correlation between the target and the features, it was found that most of the first dataset's features were correlated to heart failure, which gave the machine learning models the ability to use more information during training and testing procedure. On the other hand, only seven features from the second dataset's features showed a correlation with heart failure, which resulted in having less information for the models to use for training and testing procedures, and that could be one of the results that caused the decrease in prediction accuracy that was given from the second dataset.

Moreover, it was also found that the kind of features could affect the importance of the features to the machine learning models, and this was shown by the feature importance distribution map, which showed that the importance of a feature to one model could be differ from the other models feature importance and that means that the type of the features is important when it comes to the machine learning models and could affect their prediction accuracies. The findings that are related to the feature importance and how they affect the machine learning models' prediction accuracies answer the last question asked for this project and showed that it is crucial to carefully choose the features and the length of the dataset when studying it because they do affect the prediction accuracy.

5.2 Conclusion

To sum up, this project was made to evaluate two datasets, which were used to develop machine-learning models to predict heart failure. The main aim of this project was to enhance the early prediction of heart failure, which will allow healthcare institutions to be aware of patients who are at risk of having heart failure and to provide them with suitable treatment. Moreover, before using the machine learning models, it was essential to analyze and visualize both datasets to ensure that the datasets are prepared to be used by the machine learning models. By analyzing both datasets and

visualizing them, it was possible to have an idea of which of the features in both datasets were more correlated to heart failure; from there, the correlated features to heart failure were studied one by one and were visualized with respect to heart failure to observe how these features would lead to heart failure.

In addition, the features related to heart failure were used in some methods to train and test the machine learning models, and these methods were the splitting method, cross-validation method, and the principle component method. The first method, which is the splitting method is simply used the features of both datasets as a training set for the models, and the target for both datasets was used as a testing set to test the machine learning models. The second method, cross-validation, was used after the splitting method to enhance the prediction accuracies of the models since this method is known for its accurate calculation of the performance of models that is caused by the usage of all points in the dataset. The last method, which is principal component analysis, was also used to enhance the performance of the machine learning models. As a result, all the methods with all the models used gave reliable and acceptable prediction accuracies for both datasets; however, the first dataset, which included 49 features and 1177 patients' records, gave higher accuracies than the second dataset, included 12 features and 918 of patients' records. In addition, these results showed that the length, the amount, and the type of features included in the dataset could affect the prediction accuracy results.

It is crucial to mention the limitations that were faced during this study, which caused some trouble in having a more accurate prediction for heart failure. These limitations included some missing values from the first dataset, a shortness of length for the second dataset, a limited number of features correlated to heart failure for the second dataset, and the over-fitting and loss in information that happened when training and testing machine-learning models.

In general, the study gave excellent and reliable results and would be helpful for healthcare institutions in predicting heart failure and could also be useful for future studies.

5.3 Recommendation

The results of this study gave a valuable understanding of how to enhance the performance of the machine learning models using different methods to predict heart failure. However, during the study, it was obvious that there were some limitations

that could not be solved during the prediction process, which should be taken into consideration in future research. Firstly, it is essential to focus on the dataset's values since the missing values would affect the machine learning models' performance even if they were filled in with the fill-in techniques. The fill-in techniques in machine learning could prevent clashes caused when training and testing the machine learning modes; however, it would not prevent the bias for model predictions. The dataset length is the second thing to consider in future research since it has an essential effect on the machine learning models' predictions. It is better to consider studying large datasets, which would give machine learning models the ability to train from more examples. In addition, it would prevent the models from taking the underlying patterns, which means it prevents the over-fitting that might appear during the training part.

Generally, the machine learning techniques of filling in the missing values are considered an acceptable solution to prevent the models from facing problems like bias, loss of information, and many other issues that could affect the prediction of the models, yet having a dataset with no missing values would give better and reliable prediction accuracies from machine learning models. Lastly, it is always better to keep performing other machine learning techniques to enhance their predictions. Although multiple methods and techniques were used in this study to enhance the machine learning models' predictions, it is essential to try other new methods and techniques since machine learning is a vast field that grows massively day-by-day.

REFERENCES

- Absar, N., Das, E. K., Shoma, S. N., Khandaker, M. U., Miraz, M. H., Faruque, M. R., . . . Pathan, R. K. (2022, June 18). The Efficacy of Machine-Learning-Supported SMART System for Heart Disease Prediction. *Healthcare*, 10(6), 1137. doi:10.3390/healthcare10061137
- Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., . . . Yagil, A. (2019, November 12). Improving Risk Prediction in Heart Failure Using Machine Learning. *European Journal of Heart Failure*, 22(1), 139-147. doi:10.1002/ejhf.1628
- American College of Cardiology*. (2022, January 8). Retrieved from ACC Maintenance:
<https://www.cardiosmart.org/topics/angina#:~:text=Angina%20may%20feel%20like%20pressure,or%20discomfort%20in%20your%20chest>.
- Angraal, S., Mortazavi, B. J., Gupta, A., Khera, R., Ahmad, T., Desai, N. R., . . . Krumholz, H. M. (2020, January). Machine Learning Prediction of Mortality and Hospitalization in Heart Failure with Preserved Ejection Fraction. *JACC: Heart Failure*, 8(1), 12-21. doi:10.1016/j.jchf.2019.06.013
- Anion Gap Blood Test: Medlineplus Medical Test*. (2022, April 7). Retrieved from MedlinePlus : U.S. National Library of Medicine: <https://medlineplus.gov/lab-tests/anion-gap-blood-test/#:~:text=The%20anion%20gap%20measures%20the,blood%20isn't%20acidic%20enough>.
- Ansari, M. F., Alankar, B., & Kaur, H. (2020, July 24). A Prediction of Heart Disease Using Machine Learning Algorithms. *Advances in Intelligent Systems and Computing*, 497-504. doi:10.1007/978-3-030-51859-2_45
- Atrial Fibrillation*. (2024, March 8). Retrieved from Mayo Clinic : Mayo Foundation for Medical Education and Research: <https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/symptoms-causes/syc-20350624>
- Atypical Angina (concept ID: C0741026) - MedGen - NCBI*. (n.d.). Retrieved from National Center for Biotechnology Information. U.S. National Library of Medicine:

[https://www.ncbi.nlm.nih.gov/medgen/149267#:~:text=Definition,from%20N
CI%5D](https://www.ncbi.nlm.nih.gov/medgen/149267#:~:text=Definition,from%20NCI%5D)

- Barhoom, A. M., Almasri, A., Abu-Nasser, B. S., & Abu-Naser, S. S. (2022, April). Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms. *PhilPapers*, 6(4), 1-13.
- Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021, July 1). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. *Computational Intelligence and Neuroscience*, 2021, 1-11. doi:10.1155/2021/8387680
- Chandra , N. S., Shue Nee, S., Zhi Min, L., & Xin Ying, C. (2019, May 31). Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction. *International Journal of Innovative Computing*, 9(1), 39-46. doi:10.11113/ijic.v9n1.210
- Chauhan, Y. J. (2018). Cardiovascular Disease Prediction Using Classification Algorithms of Machine Learning. *International Journal of Science and Research (IJSR)*, 9(5), 194-200. doi:10.21275/SR20501193934
- Chicco, D., & Jurman, G. (2020, February 3). Machine Learning can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone. *BMC Medical Informatics and Decision Making*, 20(1), 1-16. doi:10.1186/s12911-020-1023-5
- Choudhary, G., & Narayan , S. (2020, October 9). Prediction of Heart Disease Using Machine Learning Algorithms. *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, 197-202. doi:10.1109/icstcee49637.2020.9276802
- Confusion Matrix in Machine Learning*. (2024, May 1). Retrieved from GeeksforGeeks: <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>
- Creatine Phosphokinase (CPK)*. (2019, March 27). Retrieved from Johns Hopkins Lupus Center: <https://www.hopkinslupus.org/lupus-tests/clinical-tests/creatine-phosphokinase-cpk/>

- Diabetes Tests*. (2023, February 28). Retrieved from Centers for Disease Control and Prevention: <https://www.cdc.gov/diabetes/basics/getting-tested.html#:~:text=Fasting%20Blood%20Sugar%20Test,higher%20indicates%20you%20have%20diabetes.>
- Duan, S., Li, Y., & Yang, P. (2023, July 31). Predictive Value of Blood Urea Nitrogen in Heart Failure: A Systematic Review and Meta-Analysis. *Frontiers in Cardiovascular Medicine*, 1-14.
- Edward R. Laskowski, M. (2022, October 8). *2 Easy, Accurate Ways to Measure Your Heart Rate*. Retrieved from Mayo Clinic : Mayo Foundation for Medical Education and Research: <https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979#:~:text=A%20normal%20resting%20heart%20rate%20for%20adults%20ranges%20from%2060,to%2040%20beats%20per%20minute.>
- El-Hasnony, I. M., Elzeki, O. M., Alshehri, A., & Salem, H. (2022, February 4). Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. *Sensors*, 22(3), 1184. doi:10.3390/s22031184
- Fedesoriano. (2021, September 10). *Heart Failure Prediction Dataset*. Retrieved from Kaggle: <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>
- Gillen, C. (2022, December 19). *Stable Angina*. Retrieved from StatPearls [Internet]. U.S. National Library of Medicine: <https://www.ncbi.nlm.nih.gov/books/NBK559016/#:~:text=Typical%20angina%20is%20often%20described,rather%20than%20a%20specific%20spot.>
- Gul, Z. (2023, January 19). *Silent Myocardial Ischemia*. Retrieved from StatPearls [Internet]. U.S. National Library of Medicine: <https://www.ncbi.nlm.nih.gov/books/NBK536915/#:~:text=Silent%20myocardial%20ischemia%20is%20a,%2C%20nausea%2C%20diaphoresis%2C%20etc.>
- Hamdaoui, H. E., Boujraf, S., Chaoui, N. E., & Maaroufi, M. (2020, September). A Clinical Support System for Prediction of Heart Disease Using Machine Learning Techniques. *2020 5th International Conference on Advanced*

Technologies for Signal and Image Processing (ATSIP), 1-5.
doi:10.1109/atsip49331.2020.9231760

Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018, December 2). A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*, 2018, 1-21. doi:10.1155/2018/3860146

Javeed, A., Khan, S. U., Ali, L., Ali, S., Imrana, Y., & Rahman, A. (2022, February 3). Machine Learning-Based Automated Diagnostic Systems Developed for Heart Failure Prediction Using Different Types of Data Modalities: A Systematic Review and Future Directions. *Computational and Mathematical Methods in Medicine*, 2022, 1-30. doi:10.1155/2022/9288452

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021, January 1). Heart Disease Prediction Using Machine Learning Algorithms. *IOP Conference Series: Materials Science and Engineering*, 1022(1), 012072. doi:10.1088/1757-899x/1022/1/012072

Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S. (2019, October). Prediction of Coronary Heart Disease Using Supervised Machine Learning Algorithms. *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 367-372. doi:10.1109/tencon.2019.8929434

Lakshmanarao, A., Swathi, Y., & Sri Sai Sundareswar, P. (2019, November). Machine Learning Techniques for Heart. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 8(11), 374-377.

Lorenzoni, G., Sabato, S. S., Lanera, C., Bottigliengo, D., Minto, C., Ocagli, H., . . . Pisanò, F. (2019, August 24). Comparison of Machine Learning Techniques for Prediction of Hospitalization in Heart Failure Patients. *Journal of Clinical Medicine*, 8(9), 1298. doi:10.3390/jcm8091298

Lutimath, N. M., C. C., & Pol, B. S. (2019, October 11). Prediction of Heart Disease Using Machine Learning. *International Journal of Recent Technology and Engineering*, 8(2S10), 474-477. doi:10.35940/ijrte.b1081.0982s1019

- Madjid, M., & Fatemi, O. (2013). Components of the Complete Blood Count as Risk Predictors for Coronary Heart Disease: In-depth Review and Update. *Texas Heart Institute Journal*, 40(1), 17-29.
- Malik, A., Brito, D., Vaqar, S., & Chhabra, L. (2023, November 5). *Congestive Heart Failure*. Retrieved from National Library of Medicine. National Center for Biotechnology Information: <https://www.ncbi.nlm.nih.gov/books/NBK430873/>
- Mortazavi, B. J., Downing, N. S., Bucholz, E. M., Dharmarajan, K., Manhapra, A., Li, S.-X., . . . Krumholz, H. M. (2016). Analysis of Machine Learning Techniques for Heart Failure Readmissions. *Circulation: Cardiovascular Quality and Outcomes*, 9(6), 629-640. doi:10.1161/circoutcomes.116.003039
- Nazrul Anuar, N., Hafifah, A. H., Mohd Zubir, S., Noraidatulakma, A., Rosmina, J., Nurul Ain, M. Y., . . . Rahman, A. J. (2020, June 19). Cardiovascular Disease Prediction from Electrocardiogram by Using Machine Learning. *International Journal of Online and Biomedical Engineering (iJOE)*, 16(7), 34-48. doi:10.3991/ijoe.v16i07.13569
- Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020, December 16). Cardiovascular Disease Prediction Using Machine Learning Models. *2020 IEEE Pune Section International Conference (PuneCon)*, 22-27. doi:10.1109/punecon50868.2020.9362367
- Olsen, C. R., Mentz, R. J., Anstrom, K. J., Page, D., & Patel, P. A. (2020, November). Clinical Applications of Machine Learning in the Diagnosis, Classification, and Prediction of Heart Failure. *American Heart Journal*, 229, 1-17. doi:10.1016/j.ahj.2020.07.009
- Pasha, S. N., Ramesh, D., Mohmmad, S., Harshavardhan, A., & Shabana. (2020, December 1). Cardiovascular Disease Prediction Using Deep Learning Techniques. *IOP Conference Series: Materials Science and Engineering*, 981(2), 022006. doi:10.1088/1757-899x/981/2/022006
- Princy, R. J., Parthasarathy, S., Hency Jose, P. S., Raj Lakshminarayanan, A., & Jeganathan, S. (2020). Prediction of Cardiac Disease Using Supervised Machine Learning Algorithms. *2020 4th International Conference on*

Intelligent Computing and Control Systems (ICICCS), 570-575.
doi:10.1109/iciccs48265.2020.9121169

professional, C. (2024). *Lactic Acidosis: Symptoms, Causes, Treatment & What It Is*. Retrieved from Cleveland Clinic: <https://my.clevelandclinic.org/health/diseases/25066-lactic-acidosis>

professional, C. (2024). *Low White Blood Cell Count (Leukopenia): Causes, Symptoms & Treatment*. Retrieved from Cleveland Clinic: <https://my.clevelandclinic.org/health/diseases/17706-low-white-blood-cell-count>

professional, C. (2024). *What is Blood Pressure?* Retrieved from Cleveland Clinic: <https://my.clevelandclinic.org/health/diagnostics/17649-blood-pressure>

professional, C. (n.d.). *Low Potassium Level Causes (Hypokalemia)*. Retrieved from Cleveland Clinic: <https://my.clevelandclinic.org/health/diseases/17740-low-potassium-levels-in-your-blood-hypokalemia>

professional, C. (n.d.). *Mean Corpuscular Volume (MCV) Blood Test V*. Retrieved from Cleveland Clinic: <https://my.clevelandclinic.org/health/diagnostics/24641-mcv-blood-test>

professional, C. (n.d.). *Noncardiac Chest Pain: Symptoms, Causes and Treatments*. Retrieved from Cleveland Clinic: <https://my.clevelandclinic.org/health/diseases/15851-gerd-non-cardiac-chest-pain>

Prothrombin Time Test. (2022, November 30). Retrieved from Mayo Clinic : Mayo Foundation for Medical Education and Research: <https://www.mayoclinic.org/tests-procedures/prothrombin-time/about/pac-20384661#:~:text=In%20healthy%20people%20an%20INR,in%20the%20leg%20or%20lung>.

RDW (Red Cell Distribution Width): Medlineplus Medical Test. (2022, April 4). Retrieved from MedlinePlus : U.S. National Library of Medicine: <https://medlineplus.gov/lab-tests/rdw-red-cell-distribution-width/#:~:text=What%20is%20a%20Red%20Cell,every%20cell%20in%20your%20body>.

- Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M. (2021, May 12). Heart Disease Prediction Using Machine Learning. *International Journal of Advanced Research in Science, Communication and Technology*, 5(1), 267-276. doi:10.48175/ijarsct-1131
- Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022, March 9). A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*, 2022, 1-9. doi:10.1155/2022/1410169
- Sahoo, P. K., & Jeripothula, P. (2021, January 29). Heart Failure Prediction Using Machine Learning Techniques. *SSRN Electronic Journal*. doi:10.2139/ssrn.3759562
- Segar, M. W., Jaeger, B. C., Patel, K. V., Nambi, V., Ndumele, C. E., Correa, A., . . . Pandey, A. (2021, June 15). Development and Validation of Machine Learning–Based Race-Specific Models to Predict 10-Year Risk of Heart Failure: A Multicohort Analysis. *Circulation*, 143(24), 2370-2383. doi:10.1161/circulationaha.120.053134
- Shahane, S. (2021, September 3). *In Hospital Mortality Prediction*. Retrieved from Kaggle: <https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction>
- Simon, L. (2023, September 4). *Hyperkalemia*. Retrieved from StatPearls [Internet]. U.S. National Library of Medicine: <https://www.ncbi.nlm.nih.gov/books/NBK470284/#:~:text=Hyperkalemia%20is%20defined%20as%20a,%2C%20muscle%20weakness%2C%20or%20paralysis>.
- Srivastava, K., & Choubey, D. K. (2020, May 30). Heart Disease Prediction Using Machine Learning and Data Mining. *International Journal of Recent Technology and Engineering (IJRTE)*, 9(1), 212-219. doi:10.35940/ijrte.f9199.059120
- Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2017). Heart failure: Diagnosis, Severity Estimation and Prediction of

Adverse Events Through Machine Learning Techniques. *Computational and Structural Biotechnology Journal*, 15, 26-47. doi:10.1016/j.csbj.2016.11.001

Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V., & Roth, G. A. (2022, December). The Global Burden of Cardiovascular Diseases and Risk. *Journal of the American College of Cardiology*, 80(25), 2361-2371. doi:10.1016/j.jacc.2022.11.005

Viral Hepatitis and Liver Disease. (2018, October 11). Retrieved from Official seal of the United States Department of Veterans Affairs: [https://www.hepatitis.va.gov/hcv/patient/diagnosis/labtests-prothrombin-time.asp#:~:text=A%20high%20PT%20usually%20means,tract%20\(esophagus%2C%20stomach\).](https://www.hepatitis.va.gov/hcv/patient/diagnosis/labtests-prothrombin-time.asp#:~:text=A%20high%20PT%20usually%20means,tract%20(esophagus%2C%20stomach).)

Wang, J. (2021, September 1). Heart Failure Prediction with Machine Learning: A Comparative Study. *Journal of Physics: Conference Series*, 2031(1), 012068. doi:10.1088/1742-6596/2031/1/012068

Zhou, J., Li, F., Song, Y., Fu, M., Han, X., & Ge, J. (2021). Prediction Model of in-hospital Mortality in Intensive Care Unit Patients with Heart Failure: Machine Learning-based, Retrospective Analysis of the MIMIC-III Database. *BMJ Open*, 11(7). doi:10.1136/bmjopen-2020-044779