Birzeit University
Faculty of Engineering & Technology
Department of Electrical & Computer Engineering
Machine Learning and Data Science Course Project

Heart Failure Prediction

Prepared By:
Eyab Ghifari - 1190999
Hamza Awashra - 1201619

Instructor's Name:
Dr. Yazan Abu Farha

Birzeit
Jan, 2024

Abstract

This study explores the use of machine learning techniques for the predictive analysis of cardiovascular diseases, employing a dataset of 918 patient records that feature essential clinical attributes. The primary aim is to utilize binary classification to determine the likelihood of patients being at risk of heart failure, incorporating a variety of numerical and categorical features. Our approach began with an exploratory data analysis (EDA), followed by the application of several machine learning models, including logistic regression, decision trees, random forests, and K-Nearest Neighbors (KNN). These models were rigorously evaluated based on metrics such as accuracy, precision, and recall. The research identified the most effective model, which was then analyzed for its clinical relevance and limitations. This study highlights the vital contribution of machine learning in improving diagnostic precision in cardiology.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI Artificial Intelligence

ML Machine Learning

MLP Multi-Layer Perceptron

SVM Support Vector Machine

KNN K-Nearest Neighbors

Chapter 1

Introduction and Motivation

Contents

1.1 Motivation

This study is driven by the critical challenge of combating cardiovascular diseases (CVDs), the leading cause of global mortality. With CVDs claiming nearly 17.9 million lives each year, there is an urgent need for early detection and management of heart failure. Leveraging the advancements in data science and the availability of extensive medical data, we aim to apply machine learning models to improve diagnostic accuracy in cardiology. This approach represents a significant step towards employing AI-driven solutions for timely and effective healthcare interventions in cardiovascular health.

1.2 Problem Statement

The abundance of medical data, coupled with advances in Data Science, has inspired numerous startups to tackle the task of developing predictive indicators for potential diseases. Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, claiming approximately 17.9 million lives annually, which represents 31% of all global deaths. Heart failure, a frequent outcome of CVDs, particularly affects those with the disease or at high risk due to factors like hypertension, diabetes, hyperlipidaemia, or existing conditions. Early detection and management are crucial for these individuals, and machine learning models offer significant assistance in this regard. By harnessing AI techniques, we aim to automate solutions for such prevalent health issues, allowing us to address subsequent challenges more effectively.

1.3 Methodology

Our methodology encompassed a comprehensive analysis of a cardiovascular disease dataset. Initially, we implemented a baseline model to establish a performance benchmark. Following this, we explored and fine-tuned two additional machine learning models, selected for their potential in handling binary classification tasks in medical datasets. The process involved rigorous hyperparameter tuning to optimize each model's performance. We then conducted an in-depth performance analysis of the most effective model, focusing particularly on identifying and understanding patterns in classification errors within the test set. This analysis was critical for evaluating the model's applicability and identifying areas for future improvement.

1.4 Report Outline

In this report, we delve into the critical field of healthcare analytics, particularly focusing on the prediction of heart failure using advanced machine learning techniques. This project not only demonstrates the power of data-driven approaches in medical diagnostics but also serves as a testament to the potential of machine learning in enhancing patient care. Below is an outline of the report, which details the comprehensive methodology and findings of our study.

1. **Introduction and Motivation:** This section introduces the study's objective, the problem statement, and the methodology used.
2. **Background:** Provides an overview of machine learning models and evaluation metrics applied in the project.
3. **Dataset Description:** Describes the dataset characteristics, data preprocessing, and exploratory data analysis.
4. **Experiments and Results:** Details the feature selection process and the comparative analysis of different models.
5. **Analysis:** Discusses the performance of the best model, including various metrics and validation techniques.
6. **Conclusions and Discussion:** Concludes with the project outcomes, model limitations, and potential areas for future research.

Chapter 2

Background

Contents

2.1 Overview

In this chapter, we present an overview of the machine learning models that have played a pivotal role in our project. These models, along with key evaluation metrics, are central to the assessment of model performance and the success of our project.

2.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a versatile supervised learning algorithm used for classification and regression tasks. It predicts outcomes for new data points based on the majority class (classification) or weighted average (regression) of their K nearest neighbors in the training dataset. The choice of K is crucial and impacts the algorithm's performance. KNN has applications in recommendation systems, image classification, anomaly detection, and more. However, it can be computationally intensive with large datasets and is sensitive to distance metrics and irrelevant features.

2.3 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It's known for its robustness, scalability, and ability to handle both classification and regression tasks. Random Forest works by creating a collection of decision trees during training and then averaging their predictions (for regression) or using majority voting (for classification) to make final predictions. This technique helps reduce overfitting and improve accuracy.

2.4 Logistic Regression

Logistic Regression is a popular linear classification algorithm used for binary and multi-class classification tasks. Despite its name, it is used for classification, not regression. It models the probability of an instance belonging to a particular class using a logistic function. Logistic Regression is simple to understand, computationally efficient, and often serves as a baseline model for classification problems.

2.5 Support Vector Machine (SVM)

Support Vector Machine is a powerful classification algorithm that aims to find a hyperplane that best separates data points into different classes while maximizing the margin between them. SVM is effective in both linear and nonlinear classification tasks, thanks to the use of kernel functions. It is known for its ability to handle high-dimensional data and perform well in various applications, including image classification, text classification, and more.

2.6 Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron (MLP) is a type of artificial neural network commonly used for deep learning tasks. It consists of multiple layers of interconnected neurons, including input, hidden, and output layers. MLPs are capable of approximating complex functions and are used in a wide range of applications, including image recognition, natural language processing, and speech recognition. Training MLPs often involves backpropagation and gradient descent.

2.7 Metrics for Evaluation

Evaluating the performance of machine learning models is essential to assess their effectiveness. Various metrics are used to measure how well a model's predictions align with the actual outcomes. Here are some common evaluation metrics:

2.7.1 Classification Metrics

Accuracy (ACC)

Accuracy measures the ratio of correctly predicted instances to the total number of instances. It's a commonly used metric for binary and multi-class classification. The formula for accuracy is:

$$\text{Accuracy (ACC)} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision and Recall

Precision and recall are vital for imbalanced datasets, where one class significantly outweighs the other. They are often used together to evaluate classification models. Precision is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-Score

The F1-Score combines precision and recall into a single metric, providing a balanced measure of a model's performance:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics are essential tools to assess the performance of machine learning models in classification tasks. The choice of a specific metric depends on the problem's characteristics and the desired trade-offs between precision and recall.

Chapter 3

Dataset Description

3.1 Overview

The dataset utilized in this project offers an in-depth examination of cardiovascular diseases, comprising clinical and demographic data of patients. It is designed to facilitate the prediction and understanding of cardiovascular health issues.

3.2 Characteristics and Attributes

- **Size and Composition:** The dataset includes 918 individual patient records, each encapsulating a variety of health indicators and personal characteristics.
- **Attributes:**
  - **Age**: Patient's age in years, providing insights into age-related risk factors.
  - **Sex**: Biological gender of the patient, crucial for understanding gender-specific trends in heart diseases.
  - **ChestPainType**: Categorization of chest pain (e.g., typical angina, atypical angina), indicative of underlying heart conditions.
  - **RestingBP**: Resting blood pressure in mm Hg, a vital sign for cardiovascular health.
  - **Cholesterol**: Serum cholesterol level in mm/dl, a key indicator of cardiovascular risk.
  - **FastingBS**: Fasting blood sugar level, with 1 indicating levels above 120 mg/dl.
  - **RestingECG**: Results of resting electrocardiograms, identifying potential heart rhythm irregularities.
  - **MaxHR**: Maximum heart rate achieved during stress, reflecting the heart's health and performance.
  - **ExerciseAngina**: Presence of angina (chest pain) induced by exercise.

- **Oldpeak**: ST depression in ECG readings post-exercise, an indicator of myocardial ischemia.
- **ST_Slope**: The slope of peak exercise ST segment in ECG, providing insights into cardiac function under stress.
- **HeartDisease**: Binary indicator of the presence or absence of heart disease, serving as the target variable for predictive analysis.

3.2.1 Statistical Overview and Exploratory Data Analysis

- **Descriptive Statistics:** The dataset was subjected to a thorough statistical analysis. Key measures such as mean, median, standard deviation, minimum, and maximum values for numerical attributes were computed to provide a quantitative understanding of the data.
- **Visual Exploratory Data Analysis:**
  - *Histograms and Boxplots* were utilized to visualize the distribution of key numerical variables like age, cholesterol levels, and blood pressure, offering insights into the spread and potential outliers.
  - *Count Plots* were generated for categorical variables such as sex, chest pain type, and fasting blood sugar, highlighting the frequency distribution of these attributes.
  - *Correlation Heatmap*: A heatmap was created to identify the relationships between different variables, particularly focusing on how they relate to the target variable, 'HeartDisease'.
- These visualizations, along with the statistical measures, provided a comprehensive overview of the dataset, allowing for initial insights into patterns and relationships within the data.

3.3 Exploratory Data Analysis (EDA)

3.3.1 Overview of The Dataset

The table below summarizes the features in the data-set. The data set has 918 observation (examples) with 12 columns (11 attributes and one target label).

Table 3.1: Heart Disease Data

Age	Sex	CP Type	R. BP	Chol	F. BS	R. ECG	Max HR	Ex. Ang.	Oldpeak	ST Slope	HD
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

The columns in the data-set encompass a diverse set of data types, including continuous variables (e.g., Age, RestingBP), binary indicators (e.g., FastingBS, HeartDisease), and derived metrics (e.g., Oldpeak). This diversity accommodates various data science approaches, from statistical analysis to predictive modeling, enhancing the dataset's utility for cardiovascular health research.

Table below summarizes the data-type used for each column and the Non-Null count for each column. Note that there are no null values for any of the attributes or for the label.

#	Column	Non-Null Count	Dtype
0	Age	918	int64
1	Sex	918	object
2	ChestPainType	918	object
3	RestingBP	918	int64
4	Cholesterol	918	int64
5	FastingBS	918	int64
6	RestingECG	918	object
7	MaxHR	918	int64
8	ExerciseAngina	918	object
9	Oldpeak	918	float64
10	ST_Slope	918	object
11	HeartDisease	918	int64
dtypes: float64(1), int64(6), object(5)			

Table 3.2: Data columns information

The table below shows the number of unique values for each column in the data-set.

Attribute	Value
Age	50
Sex	2
ChestPainType	4
RestingBP	67
Cholesterol	222
FastingBS	2
RestingECG	3
MaxHR	119
ExerciseAngina	2
Oldpeak	53
ST_Slope	3
HeartDisease	2

Table 3.3: Given Data

3.3.2 Distribution of Features Values in Dataset

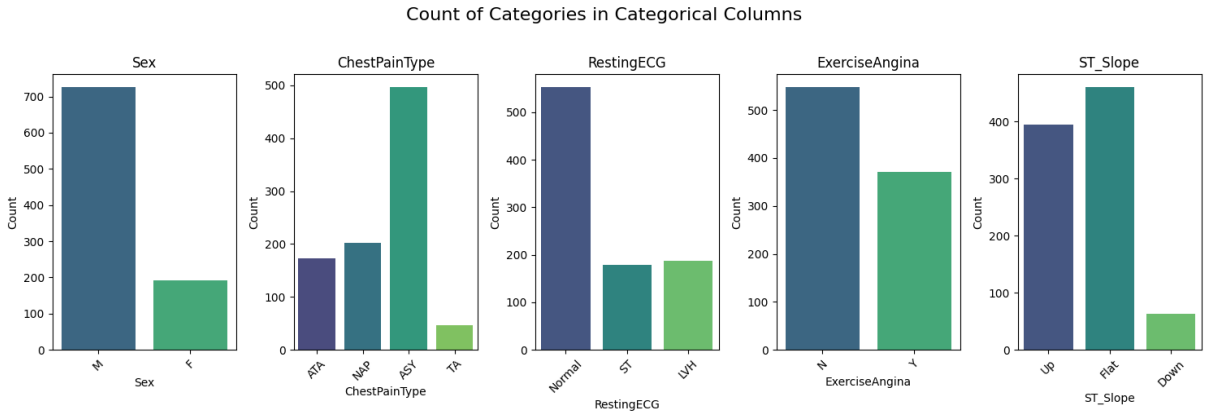

Figure 3.1: Distribution of values for categorical features

The bar charts in the figure above display the distribution of categorical variables. Males significantly outnumber females; most individuals have asymptomatic chest pain; normal ECG results are the most common; absence of exercise-induced angina is more prevalent; and a flat ST segment slope during peak exercise is observed more frequently than an upsloping or downsloping one. These visualizations indicate imbalances and prevalences in the dataset's categorical features.

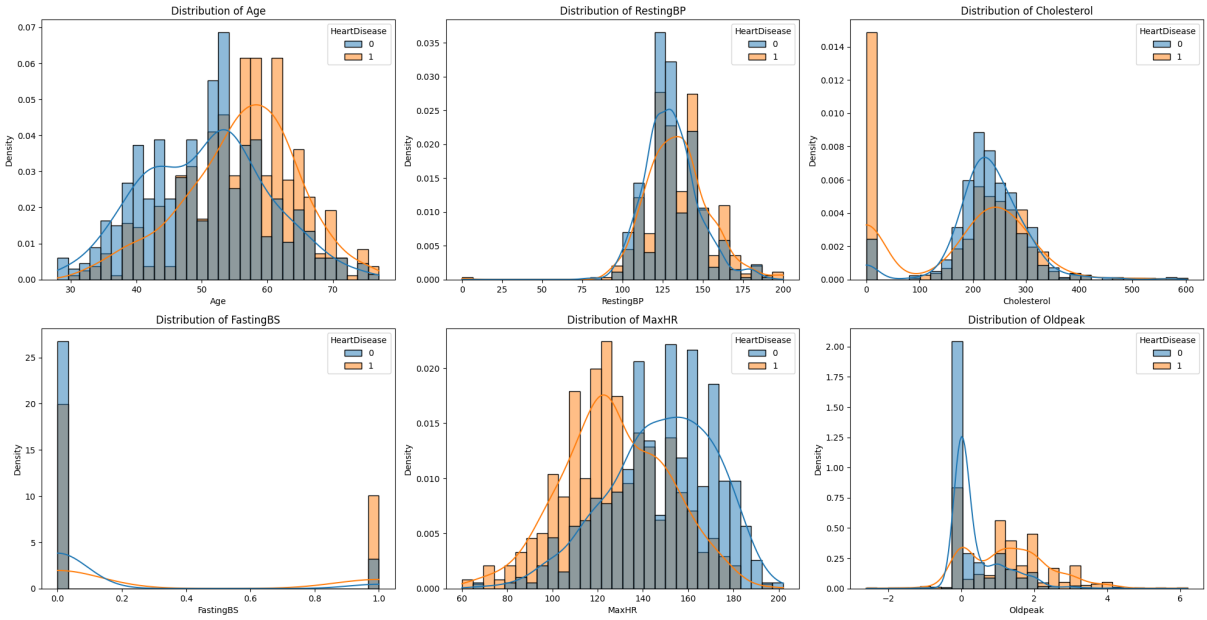

Figure 3.2: Distribution of values of the features.

The figure shown above displays the distribution of various health metrics, differentiated by the presence of heart disease. Individuals with heart disease are typically older, which is shown by the age distribution skewed towards an older demographic. Both resting blood pressure and cholesterol levels are marginally higher in heart disease patients, suggesting a subtle association with the condition. Notably, fasting blood sugar levels present a

stark contrast, with a significant skew towards higher levels in those with heart disease. Maximum heart rate reveals an inverse pattern; those without heart disease tend to have higher rates, implying that a lower maximum heart rate may be a marker of the disease. Lastly, a pronounced distribution of higher Oldpeak values, indicating more severe ST depression, is evident among those with heart disease.

3.3.3 Missing values

The analysis reveals the absence of missing values across all columns, as depicted in the table below.

Attribute	Total Missing	Percentage Missing
Age	0	0.0%
Sex	0	0.0%
ChestPainType	0	0.0%
RestingBP	0	0.0%
Cholesterol	0	0.0%
FastingBS	0	0.0%
RestingECG	0	0.0%
MaxHR	0	0.0%
ExerciseAngina	0	0.0%
Oldpeak	0	0.0%
ST_Slope	0	0.0%
HeartDisease	0	0.0%

Table 3.4: Missing Values Information

3.3.4 Skewness and Kurtosis

Skewness:

Skewness measures the asymmetry of a data distribution. It can be positive (right-skewed), negative (left-skewed), or close to zero (symmetric).

- **Symmetric:** When skewness is near zero (e.g., Age and RestingBP), the data is relatively symmetric, evenly distributed around the mean.
- **Moderately Skewed:** For moderately skewed data (e.g., Cholesterol), there's noticeable but not extreme skewness, indicating a slight bias in one direction.
- **Highly Skewed:** Highly skewed data (e.g., FastingBS and Oldpeak) have significant skewness, with one tail much longer than the other. Positive skewness means a longer right tail, while negative skewness implies a longer left tail.

The table below presents the skewness values and distribution types for each column.

Column	Skewness Value	Distribution
Age	-0.1959	symmetric
RestingBP	0.1798	symmetric
Cholesterol	-0.6101	moderately skewed
FastingBS	1.2645	highly skewed
MaxHR	-0.1444	symmetric
Oldpeak	1.0229	highly skewed
HeartDisease	-0.2151	symmetric

Table 3.5: Skewness Values and Interpretations

The skewness table indicates varying distribution shapes for health parameters: Age, RestingBP, and MaxHR are fairly symmetric (skewness near zero), while Cholesterol and Oldpeak show right-skewed distributions, indicating more extreme values above the mean. FastingBS and HeartDisease are highly skewed, reflecting binary or categorical characteristics with imbalanced class distributions. These insights highlight the diverse nature of the data distribution, guiding the choice of suitable statistical methods for further analysis.

Kurtosis:
Kurtosis measures the shape of a data distribution. It can be positive (peaked), zero (similar to a normal distribution), or negative (flatter).

- **Peaked (Leptokurtic):** Positive kurtosis (e.g., RestingBP and Oldpeak) indicates a distribution with heavier tails and a pronounced peak, signifying more extreme values and concentration around the mean.
- **Normal-Like (Mesokurtic):** When kurtosis is close to zero (e.g., Age, Cholesterol, FastingBS, and MaxHR), the data's shape resembles a normal distribution, with a moderate peak and balanced tails.
- **Flatter (Platykurtic):** Negative kurtosis (e.g., HeartDisease) suggests a flatter distribution, with lighter tails and less concentration in the center.

The table below presents the Kurtosis values and distribution types for each column.

Column	Kurtosis Value	Distribution Type
Age	-0.39	Mesokurtic
RestingBP	3.27	Leptokurtic
Cholesterol	0.12	Mesokurtic
FastingBS	-0.40	Mesokurtic
MaxHR	-0.45	Mesokurtic
Oldpeak	1.20	Leptokurtic
HeartDisease	-1.96	Leptokurtic

Table 3.6: Skewness Values and Interpretations

The table above displays kurtosis values for various medical measurements, indicating the heaviness of the tails in their distributions. Mesokurtic distributions (Age, Cholesterol,

FastingBS, MaxHR) have kurtosis values close to zero, resembling a normal distribution. Leptokurtic distributions (RestingBP, Oldpeak, HeartDisease) have positive kurtosis values, suggesting heavier tails and more outliers. In machine learning, understanding kurtosis helps in identifying distributions with extreme values, guiding data preprocessing and modeling approaches.

3.3.5 Mean, Standard-Deviation, and Quartiles

The table below shows the main, standard deviation, and other features of each column.

Table 3.7: Summary Statistics of Health Parameters

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Age	918.0	53.51	9.4	28.0	47.00	54.0	60.0	77.0
RestingBP	918.0	132.3	18.5	0.0	120.00	130.0	140.0	200.0
Cholesterol	918.0	198.7	109.38	0.0	173.25	223.0	267.0	603.0
FastingBS	918.0	0.23	0.42	0.0	0.00	0.0	0.0	1.0
MaxHR	918.0	136.80	25.46	60.0	120.00	138.0	156.0	202.0
Oldpeak	918.0	0.88	1.06	-2.6	0.00	0.6	1.5	6.2
HeartDisease	918.0	0.55	0.49	0.0	0.00	1.0	1.0	1.0

The summary statistics reveal a middle-aged cohort (mean age almost 53.5 years) with a broad age range, indicating a diverse sample. Notable findings include a slight elevation in average resting blood pressure (132.40 mmHg) and considerable variability in cholesterol levels (std = 109.38). Approximately 23 % of individuals exhibit high fasting blood sugar levels, suggesting a prevalence of glucose metabolism disorders. The data also indicates significant cardiac health risks, with 55% of participants having heart disease and a wide range of Oldpeak values, which are indicative of exercise-induced cardiac stress. The presence of physiologically implausible values (e.g., RestingBP = 0) suggests potential data quality issues that warrant further investigation.

3.3.6 Distribution of Heart Disease

The figure below depicts a pie chart that shows the percentage of positive and negative samples, in green and red, respectively.

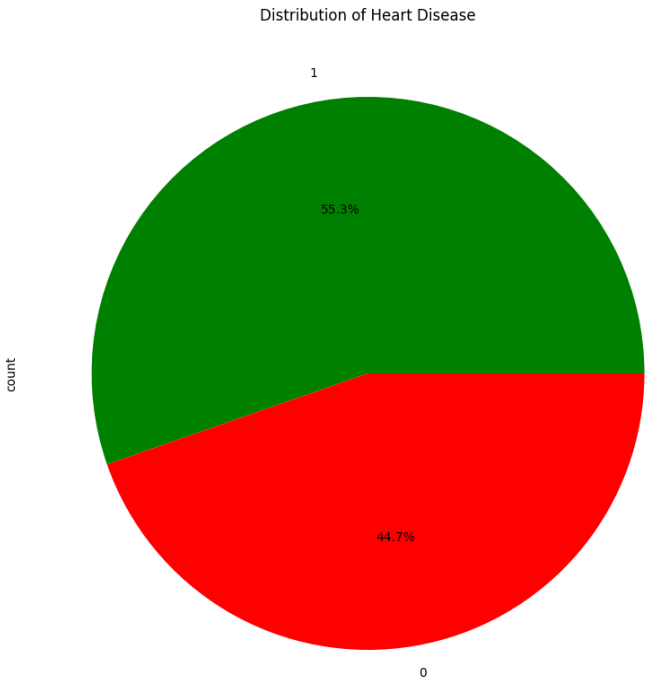

Figure 3.3: Caption for the distribution.

As we can observe from the figure above, and as was shown in a previous section, the data set is almost symmetric regarding positive and negative samples.

3.3.7 Correlation between features

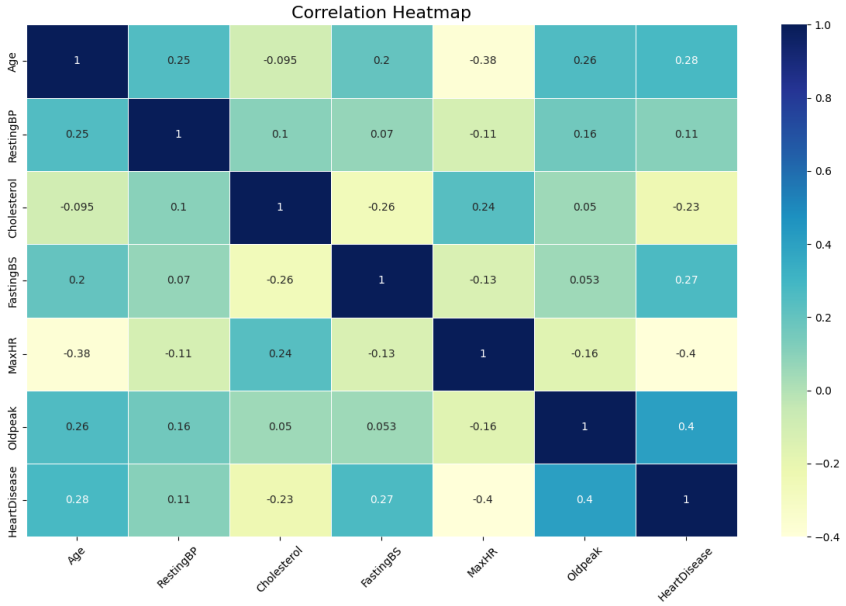

Figure 3.4: Correlation between features.

The correlation heatmap shown in the figure above indicates mostly weak to moderate correlations between various variables. Maximum Heart Rate (MaxHR) tends to de-

crease with age, and there's a moderate positive association between Fasting Blood Sugar (FastingBS) and HeartDisease, as well as between Oldpeak and HeartDisease. These relationships suggest potential predictive value for these variables in the context of heart disease.

3.4 Data Preprocessing

3.4.1 Detecting and Treating Outliers

The presence of outliers, as shown by the individual points outside the whiskers of the box plots, might necessitate outlier treatment to improve model performance. Additionally, the range and interquartile range (IQR) of the data could inform about the variance within each category and can be used to normalize or standardize the data during preprocessing.

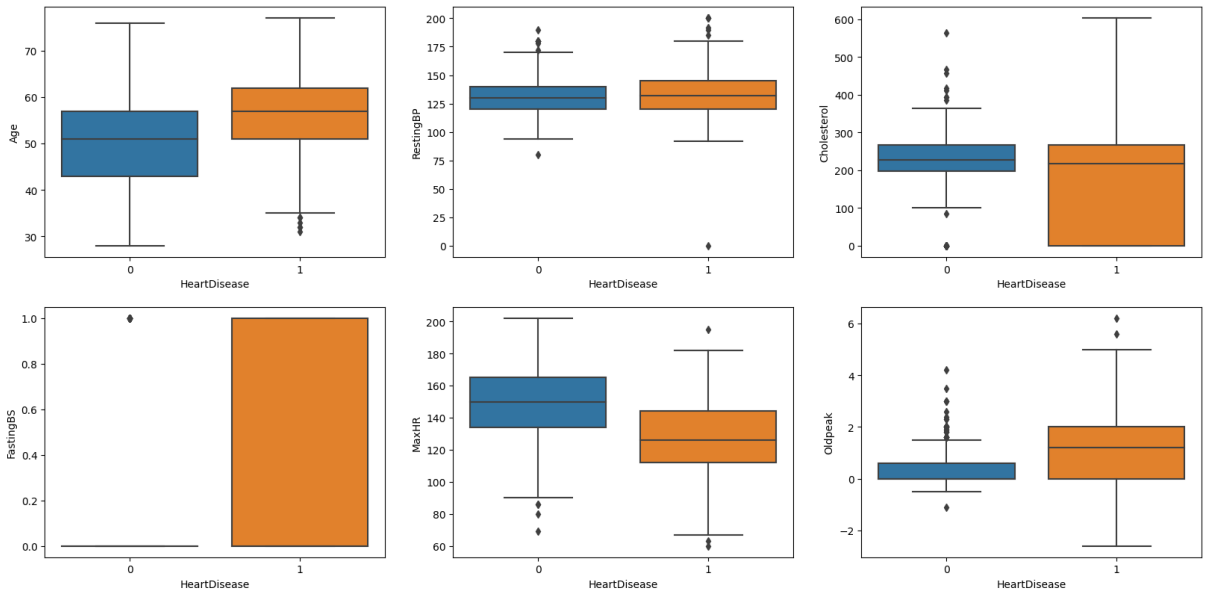

Figure 3.5: Box Plots to show outliers.

The equation below is used to replace the outliers.

$$\text{Replace} = \begin{cases} Q1 - 1.5 \times \text{IQR}, & \text{if Data} < Q1 - 1.5 \times \text{IQR} \\ Q3 + 1.5 \times \text{IQR}, & \text{if Data} > Q3 + 1.5 \times \text{IQR} \\ \text{Data}, & \text{otherwise} \end{cases} \tag{3.1}$$

where:

$Q1$ is the first quartile,
$Q3$ is the third quartile, and
IQR is the interquartile range.

The table below shows the detected outlier counts using two methods.

Measurement	IQR Outliers	Z-score Outliers
Age	0	0
RestingBP	28	8
Cholesterol	183	3
FastingBS	214	0
MaxHR	2	1
Oldpeak	16	7

Table 3.8: Outliers in Medical Measurements

The IQR method highlights extreme values based on quartile ranges, while the Z-score method identifies outliers by how many standard deviations they are from the mean. Zero outliers in Age suggest a normal distribution, whereas numerous outliers in Cholesterol and FastingBS highlight significant data dispersion, the latter likely due to its binary nature.

3.4.2 Quantification of Qualitative Data

The table below shows the new representation of the categorical data.

Patient Data					
Age	**Sex**	**ChestPainType**	**RestingBP**	**Cholesterol**	**FastingBS**
40	1	1	140	289.0	0
49	0	2	160	180.0	0
37	1	1	130	283.0	0
48	0	0	138	214.0	0
54	1	2	150	195.0	0

RestingECG	**MaxHR**	**ExerciseAngina**	**Oldpeak**	**ST_Slope**	**HeartDisease**
1	172	0	0.0	2	0
1	156	0	1.0	1	1
2	98	0	0.0	2	0
1	108	1	1.5	1	1
1	122	0	0.0	2	0

Table 3.9: Comprehensive Patient Data

As shown in the table above, the 'Sex' attribute was mapped to 1 for males and 0 for females. The heart disease is also mapped to zero (negative) and 1 (positive for heart disease).

3.4.3 Duplicate Values in the Dataset

The dataset was examined for duplicate values, and it was found that there were no duplicated rows in the dataset. The count of rows with duplicated values is zero, indicating a clean and unique dataset.

3.5 Data Quality and Integrity

The dataset is carefully curated with no missing values, ensuring a high level of data integrity for robust analysis. The data spans a diverse patient population, making it a valuable resource for understanding various facets of cardiovascular health.

3.6 Data Usage and Implications

This dataset is particularly useful for training and validating machine learning models aimed at predicting cardiovascular diseases. The diverse range of attributes allows for a multifaceted approach to understanding heart health and disease prediction.

3.7 Source and Citation

Compiled by Fedesoriano, the "Heart Failure Prediction Dataset" [1] is a publicly accessible resource hosted on Kaggle. It was retrieved on January 14, 2024, for the purpose of this project. Originating from clinical records, it offers real-world relevance for healthcare predictive modeling.

Chapter 4

Experiments and Results

Contents

4.1 Feature Selection

Using Sequential Feature Selection with a Gradient Boosting Classifier, the optimal subset of predictors for the dataset includes 'Sex', 'ChestPainType', 'Cholesterol', 'MaxHR', 'ExerciseAngina', 'Oldpeak', and 'ST-Slope'. These features yielded the best cross-validated accuracy of approximately 86.45%.

4.2 Baseline Model Evaluation

In the context of this project, the baseline model refers to a simple, initial model used as a reference point for comparing performance against more complex models developed later in the research. The k-Nearest Neighbors (kNN) algorithm has been selected as the baseline due to its simplicity, interpretability, and non-parametric nature. It requires no assumptions about the underlying data distribution, making it a flexible choice for a wide range of datasets. Additionally, kNN is intuitive to understand and implement, providing a straightforward approach to establishing initial performance metrics without the need for extensive tuning or training time.

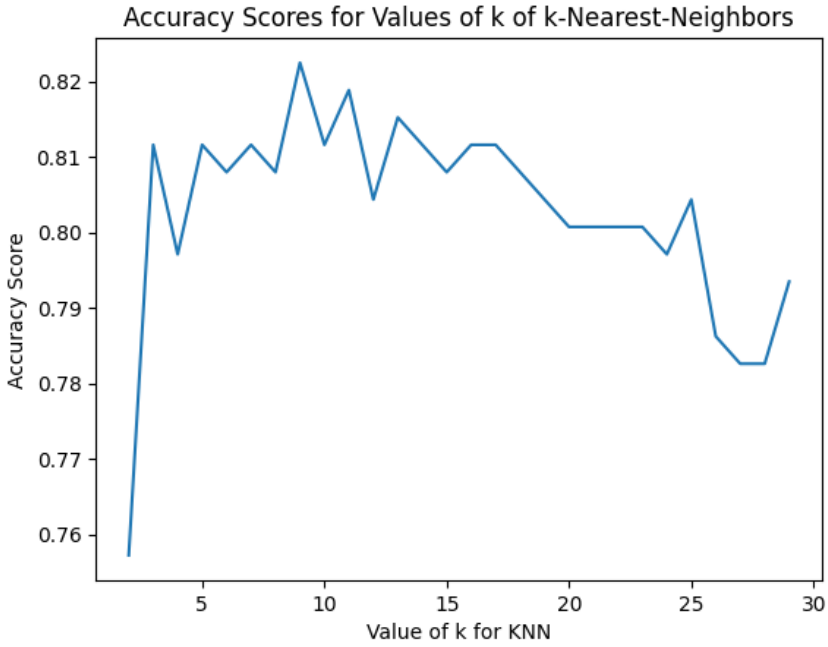

Figure 4.1: The evaluation accuracy based on different values of k in kNN

The plot of accuracy scores for different values of k in the k-Nearest-Neighbors algorithm suggests that the best value of k is **7**, as it corresponds to the peak of the highest accuracy score before the performance begins to decline. This peak indicates that a balance is achieved between underfitting and overfitting at $k = 7$, making it the optimal choice for the model given the current dataset and feature selection. Choosing this value of k would likely yield the most reliable predictions when applying the KNN classifier to new data.

The model demonstrates robust performance with an accuracy of 81.16%, indicating reliable predictions in general. The precision of 84.83% reflects a high rate of true positive

Table 4.1: Model Performance Metrics

Accuracy	Precision Score	Recall Score	F1 Score	Support
0.8116	0.8483	0.8039	0.8255	276

predictions among positive calls, while the recall of 80.39% shows the model's strength in capturing actual positive instances. The F1 score of 82.55% suggests a balanced harmony between precision and recall, marking the model's overall effectiveness in handling both false positives and false negatives. These metrics collectively signal a well-performing model for the given task.

Table 4.2: Classification Report

Class	Precision	Recall	F1-Score	Support	
Negative (0)	0.77	0.82	0.80	123	
Positive (1)	0.85	0.80	0.83	153	
Accuracy				0.81	276
Macro Avg		0.81	0.81	0.81	276
Weighted Avg		0.81	0.81	0.81	276

4.3 Advanced Models Evaluation

This section provides an evaluation of advanced models including MLP (Multilayer Perceptron), Logistic Regression, SVM (Support Vector Machine), and focuses on the Random Forest model due to its superior performance.

4.3.1 Model Selection

After thorough testing and validation, the Random Forest model was chosen due to its highest accuracy and robust performance across various metrics including precision, recall, and F1 score.

Comparative Results Analysis

4.3.1.1 MLP Classifier

- Train set Accuracy: 0.9034
- Test set Accuracy: 0.8370
- CV Accuracy: 0.8693
- Precision Score: 0.8506
- Recall Score: 0.8562
- F1 Score: 0.8534

4.3.1.2 Logistic Regression

- Train set Accuracy: 0.8567
- Test set Accuracy: 0.8116
- CV Accuracy: 0.8443
- Precision Score: 0.8483
- Recall Score: 0.8039
- F1 Score: 0.8255

4.3.1.3 SVM

- Train set Accuracy: 0.9190
- Test set Accuracy: 0.8297
- CV Accuracy: 0.8724
- Precision Score: 0.8275
- Recall Score: 0.8289
- F1 Score: 0.8281

4.3.2 Hyperparameter Tuning

In the pursuit of optimizing the Random Forest model's performance, considerable attention was devoted to hyperparameter tuning. This meticulous process encompassed the adjustment of key parameters such as the number of trees, depth of each tree, and the minimum number of samples required to split a node.

Random Forest - Chosen Model Metrics

- **Train Set Accuracy:** 0.9813
- **Test Set Accuracy:** 0.8406
- **Cross-Validation (CV) Accuracy:** 0.8817
- **Precision Score:** 0.8658
- **Recall Score:** 0.8431
- **F1 Score:** 0.8543

Parameter Grid for Hyperparameter Tuning

In the course of optimizing the Random Forest model's performance through hyperparameter tuning, a systematic exploration of parameter combinations was conducted. The parameter grid employed for this purpose was defined as follows:

- **Number of Estimators (`n_estimators`):** [10, 50, 100]
- **Criterion (`criterion`):** ['gini', 'entropy']
- **Maximum Depth (`max_depth`):** [None, 10, 20]
- **Minimum Samples Split (`min_samples_split`):** [2, 4]
- **Minimum Samples Leaf (`min_samples_leaf`):** [1, 3]

This parameter grid represents a comprehensive exploration of various configurations for the Random Forest model. It includes different numbers of estimators, criteria for splitting nodes, maximum depths of trees, and minimum sample requirements for splitting nodes and forming leaves. The hyperparameter tuning process aimed to identify the combination that maximizes the model's predictive performance.

Best Hyperparameters

The hyperparameter tuning process yielded the following optimal set of parameters:

- **Criterion:** 'gini'
- **Max Depth:** None
- **Min Samples Leaf:** 1
- **Min Samples Split:** 4
- **Number of Estimators:** 50

Grid Search Training Time

The grid search, conducted for hyperparameter optimization, incurred a training time of 41.73 seconds. This reflects the computational resources invested in achieving the optimal configuration for the Random Forest model.

Classification Report for Random Forest:

Table 4.3: Classification Metrics for Random Forest

Class	Precision	Recall	F1-Score	Support
Negative (0)	0.81	0.84	0.82	123
Positive (1)	0.87	0.84	0.85	153

The classification report provides insights into the model's performance for each class. These metrics suggest that the Random Forest model is precise in predicting positive instances and maintains a good balance in recall and F1 scores across both classes.

Chapter 5

Analysis

Contents

5.1 Performance Analysis of Best Model - Random Forest

In the pursuit of an optimal model for our dataset, various machine learning algorithms were employed and evaluated based on a set of metrics. The Random Forest model emerged as the top performer, demonstrating superior predictive capabilities. This model's performance was thoroughly assessed on different aspects, including accuracy, precision, recall, and F1 score, across training, testing, and cross-validation (CV) data sets.

5.1.1 Training Performance:

The Random Forest model exhibited exceptional performance on the training set with an accuracy of 98.13%. This high accuracy indicates that the model was able to capture the underlying patterns in the training data very well.

5.1.2 Testing Performance:

- The model achieved an accuracy of 84.06% on the test set. Although there is a drop from the training accuracy, this level of accuracy on unseen data underscores the

model's ability to generalize well.

- The precision score of 86.58% indicates the model's robustness in correctly classifying positive instances.
- The recall score of 84.31% reflects the model's proficiency in identifying all relevant instances.
- The F1 score stands at 85.43%, confirming the model's balanced performance between precision and recall.

5.1.3 Cross-Validation Performance:

The model's cross-validated accuracy is 88.17%, slightly higher than the test accuracy, reinforcing the model's consistency and capability to perform well on different subsets of the data.

5.1.4 Classification Report:

Table 5.1: Classification Metrics by Class

Class	Precision	Recall	F1-Score	Support
Negative (0)	0.81	0.84	0.82	123
Positive (1)	0.87	0.84	0.85	153

The detailed classification report provides insights into the model's performance for each class (0 and 1). These metrics suggest that the model is slightly more precise in predicting positive instances but maintains a good balance in recall and F1 scores across both classes.

5.1.5 Summary:

The Random Forest model stands out with its high degree of accuracy, precision, recall, and F1 score across the training, testing, and cross-validation datasets. It demonstrates a commendable balance between correctly predicting positive cases and covering the majority of actual positive instances. The model's ability to maintain performance consistency across different data sets highlights its robustness and reliability, making it the best choice for our predictive analysis. The precision in class predictions further assures confidence in the model's practical applicability.

Chapter 6

Conclusions and Discussion

Contents

6.1 Conclusion

In conclusion, this project on Heart Failure Prediction demonstrates the effective use of machine learning models in medical diagnostics. Through a comprehensive analysis, the Random Forest model was identified as the most efficient in predicting heart failure, owing to its accuracy and reliability. The study also underscores the significance of key evaluation metrics such as precision and recall in model assessment. While acknowledging the limitations inherent in the current models, this research paves the way for future advancements in the application of machine learning for early detection and improved management of heart-related diseases, highlighting areas that warrant further exploration and refinement.

6.2 Model Limitations

While the Random Forest model demonstrates strong performance metrics, it's crucial to recognize its limitations to comprehend when and how it can be effectively utilized. Here are some notable limitations:

1. **Model Complexity and Interpretability:** Random Forest models consist of numerous decision trees, making them complex and more challenging to interpret compared to simpler models.

2. **Performance with High-Dimensional Data:** Though capable of handling a large number of features, performance might decline in extremely high-dimensional feature spaces, particularly if there's a lot of irrelevant features or noise.

3. **Model Training Time:** Training a Random Forest can be time-consuming, especially with large datasets and many trees, which may not be ideal for time-sensitive applications.

4. **Overfitting with Noisy Data:** Despite robustness to overfitting, Random Forests can still overfit noisy datasets. Proper tuning is necessary to avoid this.

5. **Model Size:** The model can become quite large, as it consists of potentially hundreds of decision trees, leading to high memory consumption during training and prediction.

6. **Predictive Performance for Regression:** While powerful for classification tasks, Random Forests might perform less impressively with regression tasks, especially if the relationships in the data are highly non-linear or complex.

7. **Extrapolation Limitation:** Random Forests are not well-suited for extrapolation or making predictions outside the range of the training data.

Understanding these limitations is essential for applying the Random Forest model effectively and ensuring that the predictions made are reliable and applicable to the problem at hand.

Bibliography

[1] fedoriano, "Heart failure prediction dataset." https://www.kaggle.com/fedesoriano/heart-failure-prediction, Sept. 2021. Retrieved January 14, 2024.