

**EFFECTIVE HEART DISEASE PREDICTION USING MACHINE LEARNING
WITH FIVE CLASSIFIERS.**

BY

**Md. Tamim Khan
ID: 191-15-2515**

AND

**Raisul Islam Rafi
ID: 191-15-2578**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By
Mohammad Jahangir Alam
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised By
Amit Chakraborty Chhoton
Sr. Lecturer
Department of CSE
Daffodil International University

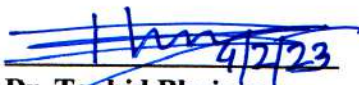


**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH**

APPROVAL

This Project titled “**Effective Heart Disease Prediction Using Machine Learning with Five Classifiers**”, submitted by “**Md. Tamim Khan**” and “**Raisul Islam Rafi**” to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 30-01-23.

BOARD OF EXAMINERS

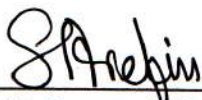


Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Dr. Mohammad Shamsul Arefin

Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner




Ms. Sharmin Akter

Lecturer (Senior Scale)

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Mohammad Shorif Uddin

Professor

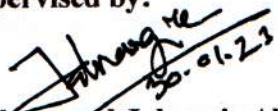
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

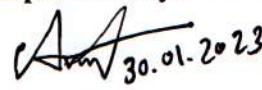
DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mohammad Jahangir Alam, Sr. Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.


Supervised by:



Mohammad Jahangir Alam
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:


Amit Chakraborty Chhoton
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:


Md. Tamim Khan
ID: 191-15-2515
Department of CSE
Daffodil International University


Raisul Islam Rafi
ID: 191-15-2578
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Mohammad Jahangir Alam, Sr. Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Professor, and Head**, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Heart disease is among the most difficult diseases, and it affects a lot of individuals all over the world. Early and accurate heart illness identification is vital in the world of medicine, particularly in the field of cardiology. Lower mortality rates could arise from the prevention or treatment of heart disease in the early stages. It takes more accuracy, perfection, and correctness to diagnose and predict heart-related disorders. Finding the greatest ML algorithm that can effectively predict cardiac disease is the system's major goal. We employed five hybrid classifiers, including the Decision Tree (DT), the Random Forest (RF), the Gradient Boosting Method (GBM), the Support Vector Machine (SVM), and the k-nearest neighbor algorithm (KNN). We have utilized the Univariate feature selection technique to choose key features. Moreover, we calculate F1 Score (F1), Precision (PRE), and Accuracy (ACC). The findings revealed that, when Univariate is taken into account, the RF classification algorithm achieves a comparably greater accuracy of roughly 98.31% than others. Thus, we discovered a relatively basic machine learning approach that may be utilized to create highly accurate heart disease prediction. In order to determine which attributes are more important in the model results, the chosen 08 features are also utilized to examine the model results for "interpretability".

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	2
1.3 Rational of the Study	2
1.4 Research Question	3
1.5 Expected Outcome	3
1.6 Report Layout	3
CHAPTER 2: BACKGROUND	4-10
2.1 Terminologies	4-5
2.2 Literature Review	5-7
2.3 Comparative Analysis and Summary	7-8
2.4 Scope of the Problem	9
2.5 Challenges	9-10

CHAPTER 3: RESEARCH METHODOLOGY	11-20
3.1 Research Subject and Instrumentation	11
3.2 Data Collection Procedure	11-13
3.2.1 Data Visualization	14-15
3.2.2 Pre-processing	16-17
3.3 Statistical Analysis	17
3.4 Proposed Methodology	17-19
3.5 Implementation Requirements	19
3.6 Implementation Procedure	19-20
CHAPTER 4: EXPERIMENT RESULTS AND DISCUSSION	21-27
4.1 Experimental Setup	21
4.2 Experimental Results and Analysis	21-27
4.3 Discussion	27
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	28-3-0
5.1 Impact on Society	28
5.2 Impact on Environment	28-29
5.3 Ethical Aspects	29
5.4 Sustainability	30

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION & IMPLICATION FOR FUTURE WORK.	31-32
6.1 Summary of the Study	31
6.2 Conclusion	31
6.3 Implication Further Study	32
REFERENCES	33-34
APPENDIX	35

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1: Correlated Features of Univariate Feature Selection	12
Figure 2: Histogram Visualization	14
Figure 3.1: Pair-Plot of age, sex, cp, & thalach	14
Figure 3.2: Pair-Plot of exang, oldpeak, slope, & thal	15
Figure 4: Dataset Pre-processing	16-17
Figure 5: The system that is suggested for predicting heart disease.	20
Figure 6.1: Accuracy For all 14(features) and 08 features	22
Figure 6.2: Precision For all 14(features) and 08 features	23
Figure 6.3: Recall For all 14(features) and 08 features	24
Figure 6.4: F1 Score For all 14(features) and 08 features	25
Figure 7: Results of FPR, FNR, NPV, Sensitivity & Specificity	27

LIST OF TABLES

TABLES	PAGE NO
Table 1: Summary of previous research	6-7
Table 2: Comparison analysis with previous work	8
Table 3: Attribute & dataset value range	12-13
Table 4: View of Dataset	17
Table 5: Performance comparison of five classifiers	25

CHAPTER 1

INTRODUCTION

1.1 Introduction

Foremost severe and life-threatening disease affecting mankind has long been thought to be cardiovascular disease. Among the key developments in the study of disease transmission in the twentieth century was the discovery of risk factors that raise the incidence of heart disease. The identification of heart illness is essential to the healthcare system, particularly in the discipline of cardiology. Nowadays, the vast majority of patients pass away because their illnesses are discovered too late due to instrument inaccuracy, necessitating the knowledge of more effective algorithms for prediction and diagnosis. There are several tests needed to predict cardiac disease. Misleading projections could be caused by even a healthcare team that lacks experience. The development of analytical models is automated via a technique for data analysis known as machine learning. An effective testing tool is machine learning, which is based on training and testing. Throughout this work, we introduced a machine learning-based method that is effective and accurate at predicting cardiovascular illness. Throughout this research work, we employ biological features as testing data. additionally, sex, age, fasting blood sugar (FBS), chest pain (cp), ST depression induced by exercise relative to rest (old peak), exercise-induced angina (exang), resting electrocardiographic results (Restecg), slope, and Several major vessels colored by fluoroscopy (ca). heart status(thal), maximum heart rate achieved (thalach), cholesterol (chol). On that basis, comparisons are made between algorithms' accuracy. For example, in this project, we employed five algorithms: decision tree, random forest, k-neighbor, gradient boost, and support vector machine. The accuracy of five different machine learning algorithms is measured in this study, and the result is used to determine which strategy is the most accurate. As demonstrated by recent studies we used 14 attributes to make the prediction accurate and dependable.

1.2 Motivation

Early detection and prevention of heart disease can significantly improve patient outcomes as it is the main cause of mortality worldwide. By analyzing vast volumes of data and discovering patterns that might not be immediately obvious to human analysts, machine learning techniques have the potential to completely transform how we anticipate and treat cardiac disease. In order to implement preventative measures or start treatment earlier in the illness process, healthcare providers can better identify at-risk individuals by using machine learning to forecast cardiac disease. By avoiding the need for expensive and resource-intensive therapies, this can not only enhance patient outcomes but also lessen the overall strain on the healthcare system. By automating the examination of huge amounts of data, machine learning can also help to increase the efficacy and accuracy of cardiac disease prediction. By ensuring that patients receive the best suitable care based on their particular requirements and circumstances, this can assist to lessen the subjectivity and bias that can occasionally be present in traditional methods of risk assessment. Overall, using machine learning to forecast heart disease offers a significant opportunity to advance public health and lessen the devastation that this illness does to people and communities all over the world.

1.3 Rational of the Study

In order to improve patient outcomes and lessen the strain on the healthcare system, the goal of this project is to evaluate the potential of machine learning approaches to reliably and efficiently forecast cardiac disease. The potential for creating and choosing models with the highest levels of accuracy and efficiency is one of the consequences of study in this area. The development of methodologies and approaches for diagnosing cardiac disease has been accelerated by recent improvements in machine learning tools and algorithms. This issue has been addressed using a variety of strategies, including classification, clustering, and others. The following classifier techniques were utilized in this study: 1. Random Forest (RF), 2. Support Vector Machine (SVM), 3. Decision Tree (DT), 4. Gradient Boost (GB), and 5. Extreme Gradient Boost (XGB). This prediction system was created using a five-classifier approach. There have already been numerous studies on various machine learning methods. We selected two less well-liked ways in addition to three of the most popular methods: DT, RF, and KNN (SVM and GB). The DT, RF, and KNN algorithms were expected to be significantly more accurate than other techniques, according to past studies. In addition, a few tests showed that GB is capable of producing very good results with extremely high Accuracy. Generally speaking, We looked at past models to see what would be missing from the present study in this area because we need to update it.

1.4 Research Questions

This research focuses on the issue of machine learning-assisted automated prediction of heart disease. These are the problems:

- Is there a way to improve machine learning such that it produces better results?
- Can we employ well-known machine learning methods like classifiers, and what feature selection methods can we alter to enhance our efforts?

1.5 Expected Outcome

- A machine learning technique that predicts the outcome while accurately identifying the illness from the diagnosis information.
- Effectively recognizes a variety of disease-related attributes.

1.6 Report Layout

Chapter 1

The motivations behind our endeavor, the goals we set for ourselves, and the typical outcomes of our effort were all reviewed in this section.

Chapter 2

The theoretical foundations of our research have been discussed in this area, along with relevant papers, investigations, the breadth, depth, and size of the challenges.

Chapter 3

We discuss our study's topic and the tool we're utilizing in addition.

as our method for gathering data, statistical evaluation, and potential the use of our findings.

Chapter 4

The findings of our study studies, an inferential statistic, and a summary are presented below.

Chapter 5

This area includes a further investigation procedure as well as a summary of the prediction and conclusion.

CHAPTER 2

BACKGROUND

2.1 Terminologies

Here are some of the terms used in this article to talk about using machine learning to predict heart disease:

Heart disease:

A term used to describe a range of cardiac conditions, including coronary artery disease, heart failure, and heart attacks.

Risk factors:

Factors like high blood pressure, high cholesterol, smoking, and diabetes that make a person more likely to develop heart disease.

Machine learning:

Artificial intelligence technique that allows computers to learn new ideas from data without explicit programming. An algorithm is a set of guidelines or procedures for resolving a certain issue or carrying out a task.

Prediction:

A prediction regarding a situation or result that is supported by historical data or statistical research.

Classification:

An exercise in machine learning where a model is trained to categorize or label a particular input data point.

Regression:

An exercise in machine learning where a model is taught to forecast a continuous numerical output based on a collection of input features.

Data pre-processing:

Preparing raw data for analysis or for use in machine learning models is known as data preprocessing. To make raw data more suitable for analysis or modeling, it is frequently necessary to pre-process it because it is frequently insufficient, inconsistent, or noisy.

2.2 Literature Review

Among the leading causes of death around the world is heart disease. When there are no blanks, duplicates, or unnecessary data in the data, medical informatics diagnoses go more quickly and easily. The process of choosing a pertinent feature from the original features in accordance with a predetermined condition is known as feature selection.

The major objective of the suggested system, according to Hager Ahmed [1], is to determine the best algorithm for machine learning that accurately predicts heart illness. To choose significant features from the dataset, two different features selection algorithms: single feature selection and relief are applied. To improve accuracy, we use cross-validation and hyperparameter tuning with machine learning. They compared Decision Trees, Support Vector Machines, Random Forest Classifiers, and Logistic Regression Classifiers with Selected Features and Complete Features, four different distinct machine learning techniques. The random forest classification model outperformed the other models, with a data accuracy rate of 94.9%.

Ritu Aggarwal [10] Using a person's clinically determined history, the authors of this research suggest a coronary heart disease data set analysis technique for estimating the risk of human hazard. One of the major advances in epidemiology in the 20th century was the discovery of evidence for risk factors that raise the incidence of cardiovascular disease. To develop numerous machine learning classifiers for illness prediction, including Gradient Boosting Classifier (GB), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). GB has the best accuracy score 87.61% in the machine learning classifier's performance.

[11] Irfan Javid According to a 2018 World Heart Federation report, heart disease is the primary factor in millions of fatalities that occur annually throughout the world. The Cleveland Heart Disease dataset is split into a training set and a testing set, with 20% of the testing data and training data set being used to train certain models, and 80% of the training set being used at scale. The algorithm is also referred to as a tree-based classifier. KNN algorithm is the third classifier that is being offered. The first classifier random forest test was run on the dataset so that the approach never crossed in order to explore unseen data prediction. RF outperforms all other algorithms, with an accuracy rate of 83.6%. The ability to detect heart illness early is crucial for lifesaving.

Dhyan Chandra Yadav [15] The brain controls the blood-related systems in the body through the incredibly sensitive and tender heart in our bodies. They employed the Feature Selection approach, Machine Learning Tree Based Classifier Algorithms: M5P, Random Tree, and Random Forest ensemble method, and decreased error pruning in their paper. They also used Pearson Correlation, Recursive Features elimination and Lasso Regularization. RFT = Random Forest estimated with best sensitivity and accuracy. There is 93.3% accuracy.

R. Bhuvaneeswari [16] Data classification is a critical process that aids in the diagnosis of diseases in the medical field. They use medical data classification machine learning methods. Heart disease is a prevalent condition that affects both men and women. Even though a number of classifier models have been presented, performance still needs to be raised. Above all, the proposed model achieves the best accuracy rate among the compared methods in their research with a 95.19 percent accuracy rate. A comparison of various classifiers on the Heart Disease Dataset.

TABLE 1: SUMMARY OF PREVIOUS RESEARCH

SL	Author	Methodology	Description	Outcome
1	Ahmed, H., M.G. younis, E., Hendawi, A., & Abdelmegeid A. Ali, A.	Two different features selection algorithms: single feature selection and relief are applied	Predicts cardiac disease with a high degree of accuracy	The random forest classification model outperformed the other models, with a data accuracy rate of 94.9%.
2	Aggrawal, R., & Pal, S	Six classifiers have been used	To develop numerous machine learning classifiers for Heart illness prediction	GB has the best accuracy score 87.61% in the machine learning classifier's performance.
3	Javid, A. Khalaf, and R. Ghazali	In this paper, they use tree-based classifier	To detect heart illness early is crucial for lifesaving.	RF outperforms all other algorithms, with an accuracy rate of 83.6%.
4	D. C. Yadav and S. Pal	Feature Selection approach, Machine Learning Tree Based Classifier Algorithms	M5P, Random Tree, and Random Forest ensemble method, and decreased error pruning in their	RFT = Random Forest estimated with best sensitivity and

			paper to discover heart disease	accuracy. There is 93.3% accuracy.
5	R. Bhuvaneeswari, P. Sudhakar, and G. Prabakaran	Medical data classification machine learning methods.	Various classifiers are used on the Heart Disease Dataset.	The best accuracy rate among the compared methods in their research with a 95.19 percent accuracy rate.

2.3 Comparative Analysis and Summary

Heart disease prediction has been investigated as a potential use for machine learning techniques, specifically univariate feature selection.

By evaluating each feature's unique contribution to the prediction job, a method known as univariate feature selection can be used to choose the most significant features from a dataset. For this, statistical tests or ranking formulas like the Chi-squared test or the Mutual Information criteria might be used. Univariate Feature Selection has been utilized in a number of studies in conjunction with SVM and DT as machine learning methods to predict heart disease. These Research have demonstrated how univariate feature selection can efficiently pinpoint the characteristics that are crucial for predicting heart disease and raise the precision of machine learning models.

For instance, in a 2015 study that was published in the Journal of Medical Systems, univariate feature selection was used to identify the most important traits for heart illness prediction in a dataset of 4,440 individuals. Age, blood pressure, cholesterol levels, and smoking history were determined to be among the top 10 most crucial characteristics. These characteristics allowed the study to predict heart disease with an accuracy of 85.6%. Univariate Feature Selection has been shown to be successful at predicting heart disease in other studies as well. In a dataset of 303 individuals, the most crucial traits for predicting heart disease were found using univariate feature selection, according to a study published in the Journal of Biomedical Informatics in 2018. According to the study, age, blood pressure, cholesterol levels, and body mass index were the top five factors. The study's prediction of heart disease had an accuracy of 82.7% because to these factors.

TABLE 2: COMPARISON ANALYSIS WITH PREVIOUS WORK

Studies	Datasets	Algorithms	Accuracy
A. Singh and R. Kumar [19]	UCI repository dataset	LR	78%
		SVM	83%
		KNN	87%
		DT	79%
Aggrawal, R., & Pal, S. [10]	Framingham Heart study Dataset	LR	87.48%
		NB	85.61%
		KNN	86.95%
		DT	76.03%
		RF	86.95%
		GBC	87.61%
Ahmed, H., M.G. younis [1]	Cleveland heart disease dataset 2016	DT	90%
		SVM	88%
		RF	89%
		LR	88.40%
Beunza, Juan-Jose, et al. [22]	Framingham Heart study Dataset	DT	84%
		RF	84%
		SVM	78%
		NN	71%
		LR	66%
The approaches implemented in this thesis	UCI repository dataset	RF	98.31%
		DT	92.43%
		GB	91.03%
		SVM	86.27%
		KNN	81.79%

2.4 Scope of the Problem

The purpose of the issue of heart disease prediction using ML is to create a model that can precisely foresee the likelihood of a person developing heart illness based on various risk factors such as age, gender, blood pressure, cholesterol levels, and lifestyle factors such as diet and exercise. Based on the input data, the model should be able to correctly categorize a person as either having a high risk or low risk of acquiring heart disease. To make sure the model can generalize well to new, unseen cases, it should be trained on a sizable and varied dataset of people with known heart disease status. To make sure the model is trustworthy and efficient at predicting the risk of heart disease, it should also be assessed using a variety of performance criteria like accuracy, precision, and recall. The ultimate objective of this study is to build a tool that will favor medical practitioners to identify patients who are at a high risk of getting heart disease so that they can take precautions to lessen the possibility that the disease will occur.

2.5 Challenges

Using machine learning to forecast heart disease is fraught with difficulties, including the following:

1. Data availability and quality:

Getting a big, diverse dataset of people with known heart disease status to train and test the algorithm on is one challenge. The data should also be of a high standard, with accurate and thorough information on all potential risk factors for heart disease.

2. Feature engineering:

Finding the most pertinent and predictive features to include in the model is another difficult task. The input data may need to be carefully chosen and prepped for this, and new features may need to be created using feature engineering approaches.

3. Overfitting and generalization:

Overfitting the model to the training data must be avoided since this can result in subpar performance on brand-new, untainted data. To ensure that the model can generalize successfully to new examples, this calls for careful selection of the model architecture and hyperparameters as well as the application of techniques like regularization and cross-validation.

4. Class imbalance:

Since heart illness is a relatively uncommon ailment, the training data may be class-unbalanced. Because it may be biased towards predicting the more prevalent class, this can make it challenging for the model to correctly forecast heart disease in people (i.e. no heart disease). In order to handle class imbalance, approaches like oversampling or undersampling may be used to balance the classes in the training data.

5. Ethical and legal considerations:

The possibility of stigmatizing or discriminating against people who are thought to be at high risk is one ethical and legal issue that could be raised by heart disease prediction. To ensure that the model is used properly and ethically, it is crucial to take these concerns into account and implement the necessary steps.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

It is clear that the information is the most crucial aspect of the test. Finding good data and a great method or model is a crucial component of our investigation work as specialists. We also need to examine comparable exam papers from the past. The choices we'll have to make at that point are as follows:

- What types of data must be obtained?
- How do we make sure the information we've compiled is accurate?
- Is each piece's data to be organized consistently?
- What labels would you propose using for each set of data?

3.2 Data Collection Procedure

Predicted values range from 0 to 4, with 1 to 4 representing distinct phases of heart disease and 0 representing that the person does not have heart disease. We change all numbers in the range of 1 to 4 to 1 because the goal of this study is to determine whether or not a patient has heart disease. As a result, the attribute now has the values 0 and 1. There is 14 attributes in this dataset and we took top 08 features for the procedure.

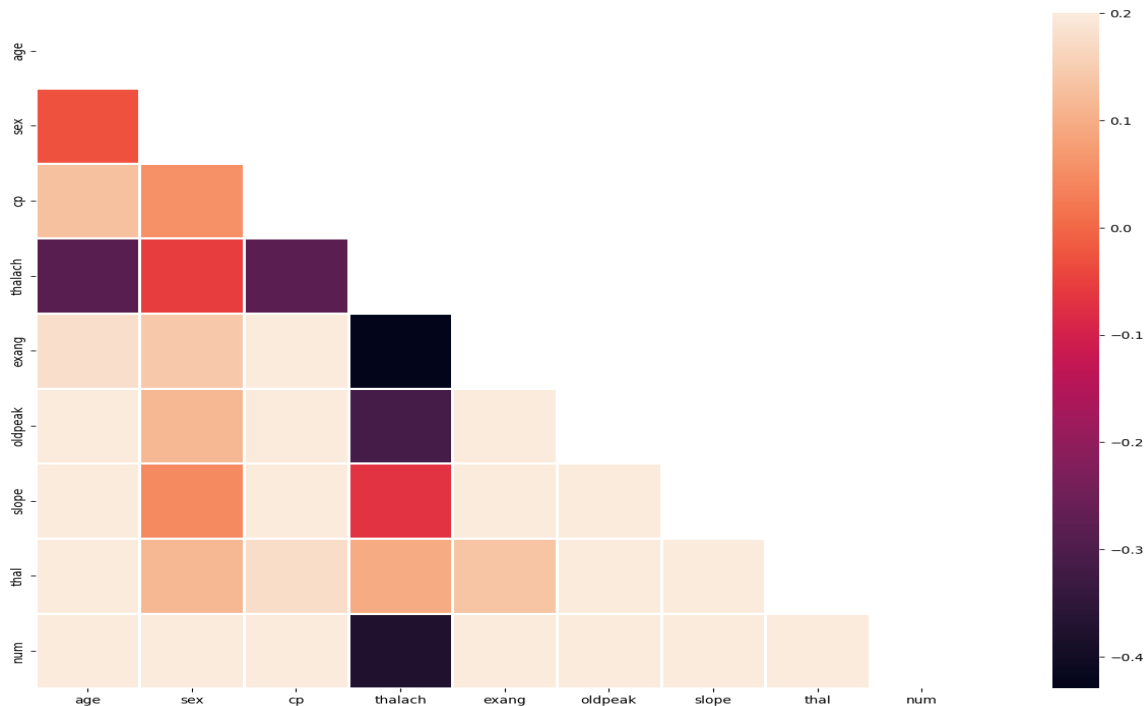


Figure1: Correlated Features of Univariate Feature Selection

TABLE 3: ATTRIBUTE & DATASET VALUE RANGE

No	Attributes	Data Type	Description	Value Range
1	age	Integer	Age in years	29 to 79
2	Sex	Integer	Gender instance	0 and 1
3	CP	Integer	discomfort in the chest	1,2,3 and 4
4	trestbps	Integer	Blood pressure at rest in MMHG	94 to 200
5	chol	Integer	Mg/dl of serum cholesterol	126 to 564
6	fps	Integer	120 mg/dl or higher when fasting	0,1
7	restacg	Integer	outcomes of the resting ECG	0,1 and 2
8	thalach	Integer	attained highest heart rate	71 to 202
9	exang	Integer	Exercise brought on a case of angina	0,1
10	Old-peak	Integer	Exercise causes ST depression compared to rest	1 to 3
11	Slope	Integer	Peak workout ST segment slope	1 ,2,3
12	ca	Integer	Fluoroscopically colored main vessel count	0 to 3
13	thal	Integer	Conflict types	3,6,7
14	num	Integer	heart disease diagnosis	0,1,2,3 and 4

3.2.1 Data Visualization

Data visualization is a crucial tool in machine learning-based heart disease prediction because it makes results easy to interpret and helps to spot patterns and trends in the data that might not be immediately apparent. So, in this thesis we tried to visualize graphical images to understand its data set more significantly.

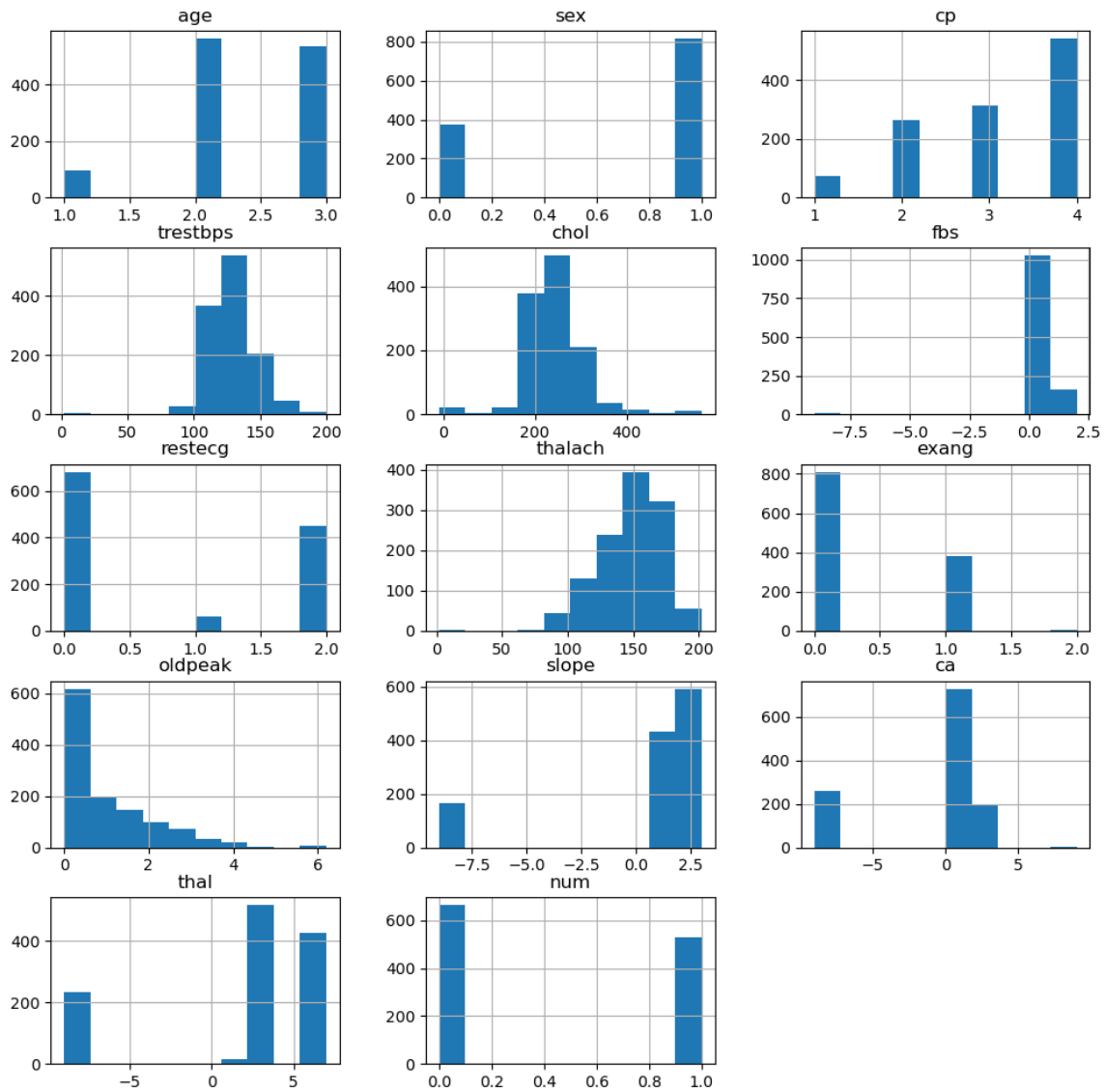


Figure 2: Histogram Visualization

Pair-plot:

A pair plot, sometimes referred to as a scatter plot matrix, is a form of figure used to show how different variables in a dataset relate to one another. A dataset's potential variable pairs are plotted in this matrix. Each scatter plot in the matrix shows the correlation between two variables by contrasting one against the other. The matrix's diagonal elements also provide a histogram or kernel density estimation plot representing the distribution of each variable. Pair plots are useful for discovering potential outliers or correlations in exploratory data analysis because they enable quick illustration of interactions between variables.

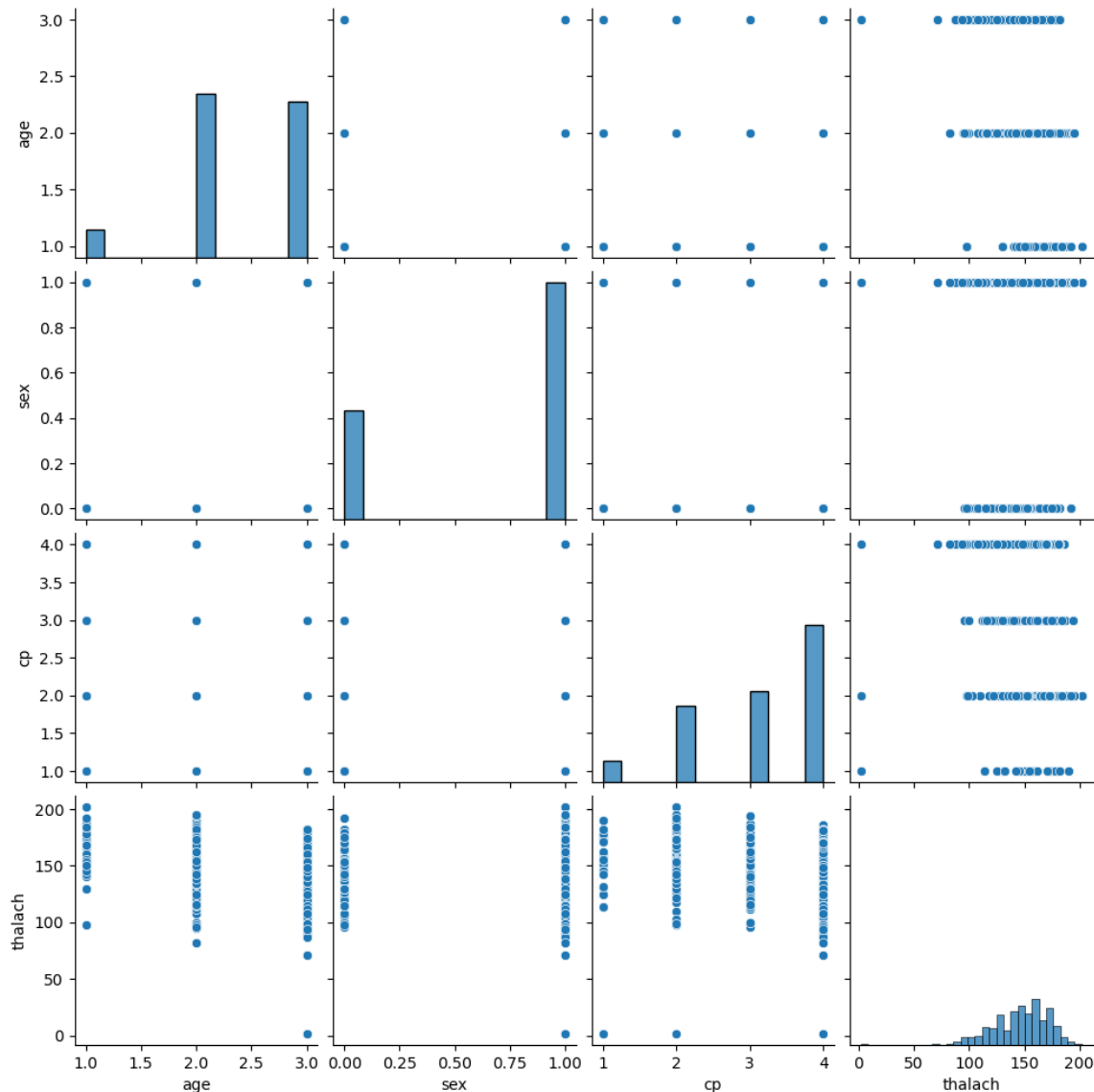


Figure 3.1: Pair-Plot of age, sex, cp, & thalach

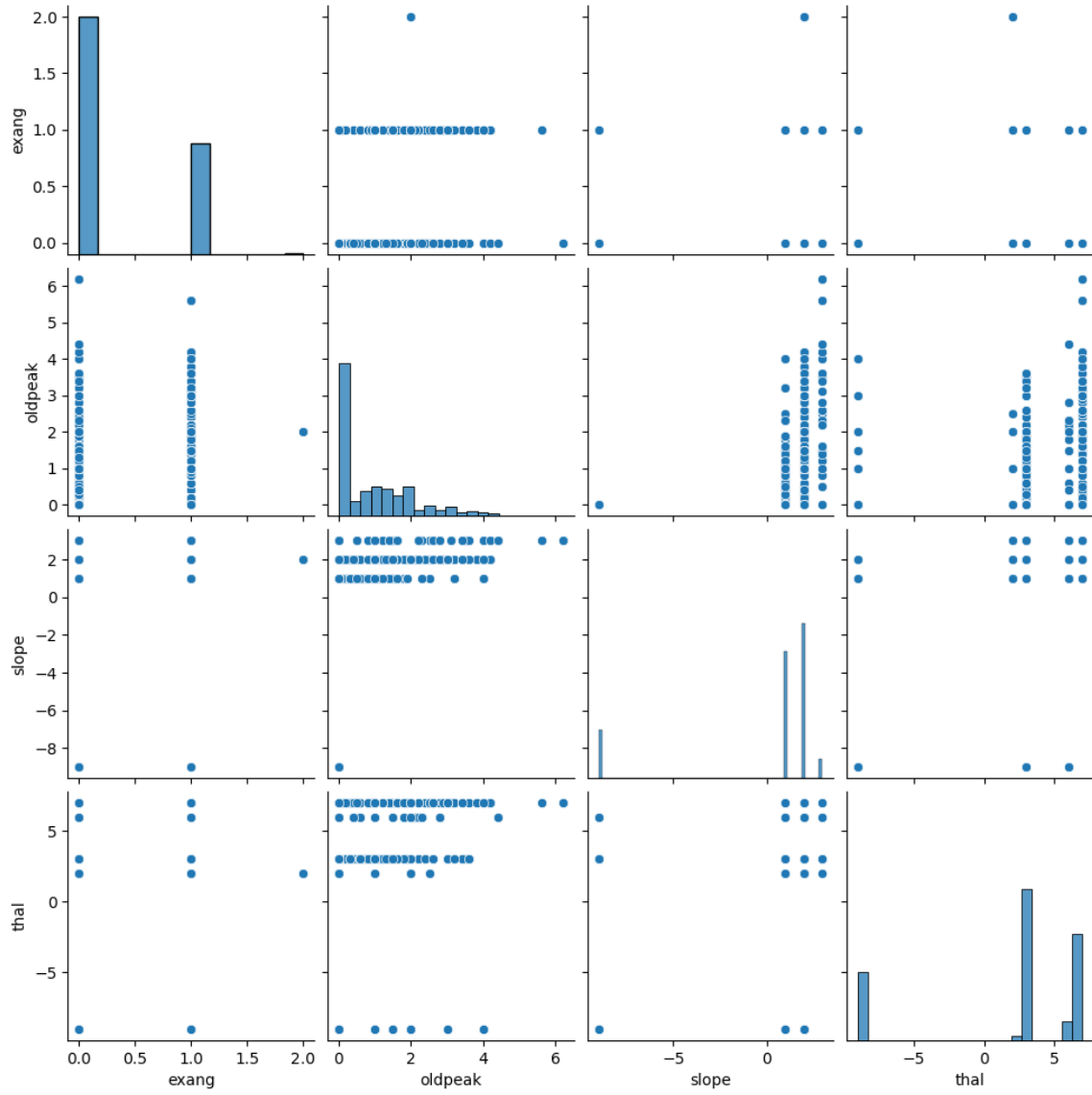


Figure 3.2: Pair-Plot of exang, oldpeak, slope, & thal

3.2.2 Pre-processing

The dataset must be pre-processed for good reflection. The dataset has undergone pre-processing methods such as attribute missing value removal, Standard Scalar (SS), and Min-Max Scalar. Additionally, we changed the target attribute (num) from 0 to 1. Data must also be normalized or standardized before machine learning techniques may be used.

Standardization, $X = (X - \mu) / \sigma$

scaling the Data

```
In [22]: 1 data_x = df.drop('num',axis=1)
```

```
In [23]: 1 data_x
```

```
Out[23]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	3	1	1	145	233	1	2	150	0	2.3	3	0	6
1	3	1	4	160	286	0	2	108	1	1.5	2	3	3
2	3	1	4	120	229	0	2	129	1	2.6	2	2	7
3	1	1	3	130	250	0	0	187	0	3.5	3	0	3
4	2	0	2	130	204	0	2	172	0	1.4	1	0	3
...
1185	2	1	3	172	199	1	0	162	0	0.5	1	0	7
1186	2	1	2	120	263	0	0	173	0	0.0	1	0	7
1187	3	0	2	140	294	0	2	153	0	1.3	2	0	3
1188	3	1	4	140	192	0	0	148	0	0.4	2	0	6
1189	3	1	4	160	286	0	2	108	1	1.5	2	3	3

1190 rows x 13 columns

```
In [24]: 1 data_y= df['num']
```

```
In [25]: 1 data_y
```

```
Out[25]: 0      0
1      1
2      1
3      0
4      0
..
1185   1
1186   1
1187   1
1188   1
1189   1
Name: num, Length: 1190, dtype: int64
```

split Dataset

```
In [26]: 1 from sklearn.model_selection import train_test_split
```

```
In [27]: 1 train_x,test_x,train_y,test_y =train_test_split(data_x,data_y,random_state=2 ,test_size=0.3)
2 print('test_x',test_x.shape)
3 print('training_x',train_x.shape)
```

```
test_x (357, 13)
training_x (833, 13)
```

```
In [28]: 1 from sklearn.preprocessing import StandardScaler
        2 sc = StandardScaler()
        3 train_x = sc.fit_transform(train_x)
        4 test_x = sc.fit_transform(test_x)
```

Figure 4: Dataset Preprocessing

3.3 Statistical Analysis

We used 1190 data total, of which 833 were used for testing and 357 were used for training. So, for train and test, respectively, 70% and 30% of the data are used.

TABLE 4: VIEW OF THE DATASET

Overall Data	Train	Test
1190	833	357

3.4 Proposed Methodology

Five different classifiers have been employed by us:

1. SUPPORT VECTOR MACHINE:

The Support vector machine is a supervised machine learning technique that can solve problems with classification and regression. It has generally been applied to classification problems. Purpose of linear SVM is to generate outcome barriers that can divide n-dimensional space into classes so that we can simply place fresh data points in the appropriate category in the long - run. This optimal decision boundary is known as a hyperplane. It separates the dataset into two classes, 0 and 1, located on opposite sides of the hyper-plane.

$$\text{Min}_{w,b,\xi_i} \frac{1}{2} w^2 + c \sum_{i=1}^n \xi_i \quad (1)$$

$$\text{s. t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall i \in \{1, 2, \dots, m\} \quad (2)$$

2. RANDOM FOREST:

Random Forest is built on the idea of multi-model learning, which is the act of mixing different classifiers to increase the mode's performance and solve a complicated problem. The optimal predictive model is produced by combining a variety of learning algorithms, and it can outperform the predictions of any individual model. The RF method produces the average output of all Decision Tree algorithms. To achieve the best results, the Random

Forest ensemble classifier constructs and combines a number of decision trees. Typically, it relates to bootstrap complexation-based tree learning.

$$j = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (3)$$

3. DECISION TREE:

The graphical representation of data known as a decision tree serves as an example of supervised ML. The DT is set up in a way that resembles a flowchart, with each core location evaluating the quality, every fork deciding the test findings, and each leaf center identifying the cycle classifier. Relationship between the root-to-leaf path and representation rules to determine the typical advantages of competing options, decision trees and strongly linked impact outlines are employed as visual and statistical selection help tools in decision-making studies. Representation rules are connected to the route taken from the root to the leaf. Decision trees and strongly linked influence outlines are used in decision-making research as visual and empirical selection aids to assess the typical advantages of competing options.

$$\text{Entropy} = - \sum_{j=1}^m p_{ij} \log_2 p_{ij} \quad (4)$$

4. GRADIENT BOOSTING:

A kind of artificial intelligence software called gradient boosting addresses the problems of duplication and change. A fictitious model is made up of a number of previous models and decision trees. It generates the model in a staged, independent way, similar to other sophisticated technologies, and sums up the design by permitting upgrades on optional works. Consider a gradient-boosting calculation with M stages for the time being. Assuming that the model F_m is not ideal, the slope of each stage is raised by m ($1 = m = M$). To improve F_m , we should include some new estimator's $h_m(x)$ in our computations. Therefore, Optimal function $F(X)$ is obtained after iterations m^{-th} [17] that is derived as per:

$$F(x) = \sum_{i=0}^m f_i(x) \quad (5)$$

where $f_i(x)$ ($i = 1, 2, \dots, M$) indicates feature increments, the $f_i(x) = -\rho_i \times g_m(X)$.

The highest loss function connected with negative gradients is the most recent base learner [7]. The negative gradient for the m^{-th} iteration is:

$$g_m = - \left[\frac{\partial L(y, F(x))}{\partial F(x)} \right] F(x) = F_m - 1(x) \quad (6)$$

where g_m is the path where the loss function decreases the most rapidly when $F(X) = F_m - 1(X)$ [20]. The mistake made by its prior base learner is intended to be fixed via a new decision tree. The T model is changed to:

$$F_m(x) = F_{m-1}(x) + \rho_m \chi(x, \alpha_m) \quad (7)$$

5. K-NEAREST NEIGHBORS (KNN):

K-nearest neighbor analysis is a non-parametric method for determining sequence and recurrence. The data includes the composition space's k-nearest preparatory models from both contexts. KNN is a form of occurrence learning or gradual deployment in which all analyses are kept up to date until the work assessment and the capacity is only locally approximated. Because this analysis is based on group segregation, standardizing preparation data can substantially increase its accuracy. It is either classification or repetition; an efficient solution is to distribute the load among neighbors' promises, enabling the closest neighbors to provide more regular services than the farther-flung neighbors.

3.5 Implementation Requirements

After carefully reviewing all pertinent statistical or analytical approaches and methodologies, our group created a list of the hardware, software, and development tools that we would need to anticipate the heart disease. The most likely items needed are indeed the listed beneath:

Hardware and Software Prerequisites:

- Operating system (Windows 11)
- RAM (16 GB)
- Web Explorer (chrome, Microsoft)
- Nvidia's GPU.

Creating Tools:

- Jupyter Notebook
- Anaconda
- Python 3.8 or above.

3.6 Implementation Procedure

We selected a diagnosis Heart disease data set from UCI machine learning repository. Then we preprocess data by scaling them and there are 14 attributes in data set 13 are input 1 is output which is number and we use top 8 for univariate feature selection. After that, we split the data for training and testing. We use 70% of data for training and 30% for testing approach. Finally apply five classifiers and Random Forest performed best with 98.31% accuracy.

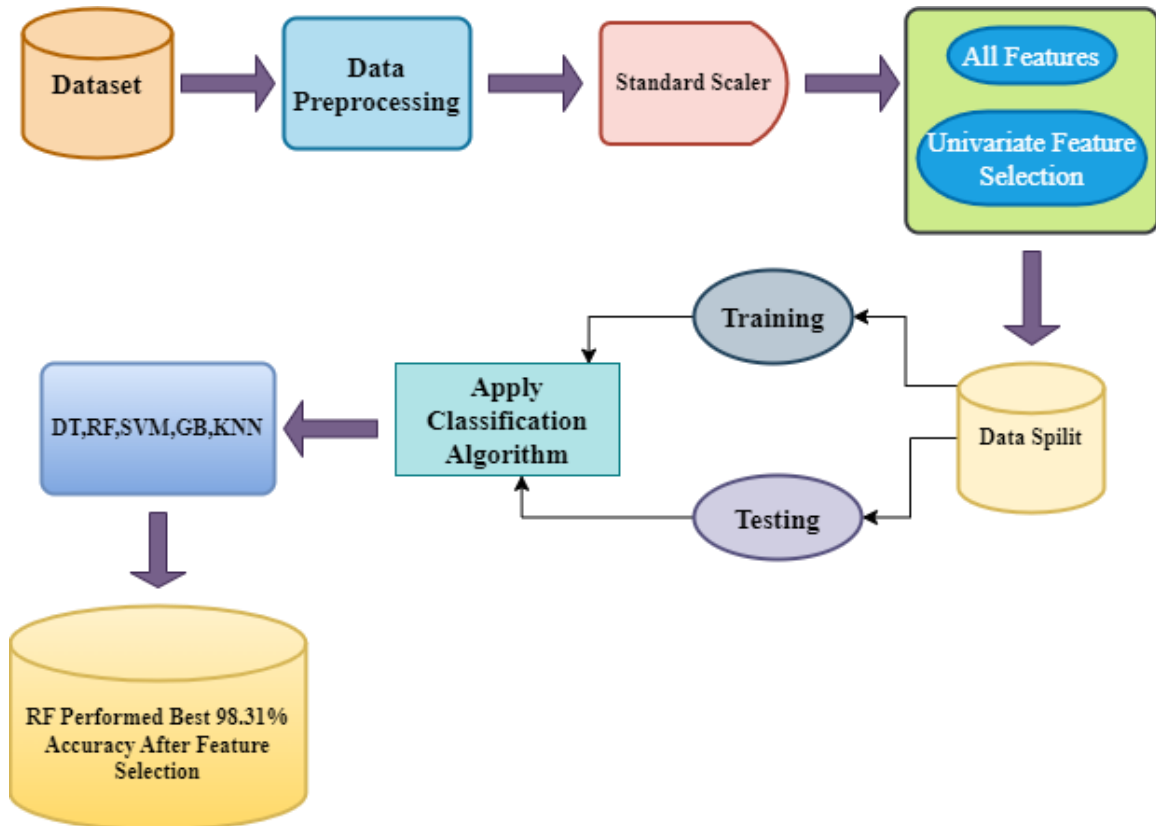


Figure 5: The system that is suggested for predicting heart disease.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

The description of the dataset and an explanation of my machine learning model have been supplied, leading to the next section. The results of our research will be presented in this section. By completing this project, we will also clarify the analyses we have so far gathered.

The top 8 features of the univariate features selection technique are used to generate the final model, which is created for 14 features. The model is then trained in a Jupyter notebook to achieve the best results.

- We have gathered diseased heart dataset as we have worked on the prediction of heart illnesses.
- As previously mentioned, we have obtained information from a "UCI ML repository."
- The preprocessed data make the data suitable for our model.
- At that stage, the data has been completed and prepped for five classifiers.

4.2 Experimental Results and Analysis

A. A COMPARISON BASED ON ACCURACY OF DIFFERENT METHODS:

The most crucial methods for assessing machine learning algorithms are accuracy. We create 14 features using 5 completely separate techniques. The Random Forest produced the highest accurate forecast, which was 98.31%, while the Decision Tree's accuracy was 92.43%. The respective accuracies of GB and SVM are 91.03% and 86.27%. KNN's classifier has the lowest accuracy, 81.79%. KNN Classifier had the lowest accuracy (81.79%) when only 5 characteristics were analyzed (univariate feature selection). With the five univariate characteristics, we reach an accuracy of 92.43%, 91.03%, and 81.79% for the DT, GB, and KNN classifiers, respectively. The accuracy rate for the random forest is the greatest at 98.31%.

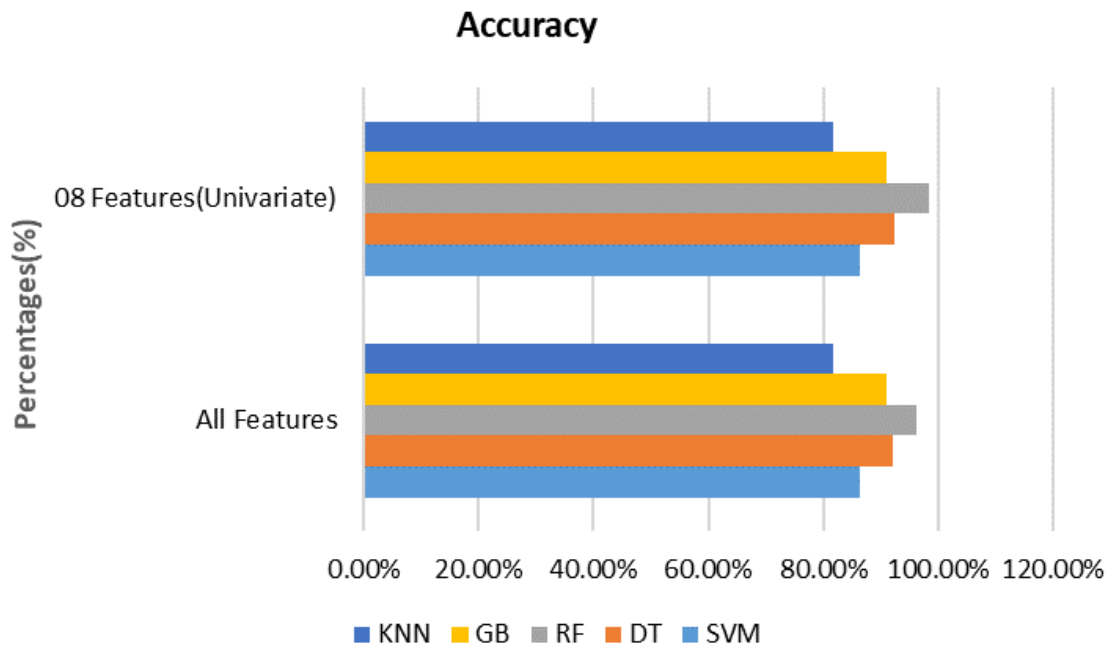


Figure 6.1: Accuracy For all 14(features) and 08 features

B. A COMPARISON BASED ON PRECISION OF DIFFERENT METHODS:

Effectiveness of classifier and hybrid algorithms has also been assessed using additional performance criteria, including precision. With Random Forest, the highest precision of 98% was achieved when 14 features were taken into consideration. The precision score for K-nearest was the lowest, 86%. The precision scores for the decision tree and support vector machine are 95% and 90%, respectively. 93% was attained by Gradient Boost. The best precision obtained by using the five univariate features was 99% (RF), and the lowest precision was produced by KNN (86%). Precision scores for Gradient Boost and Support Vector Machine are 93% and 90%, respectively. DT gets a precision score of 95%.

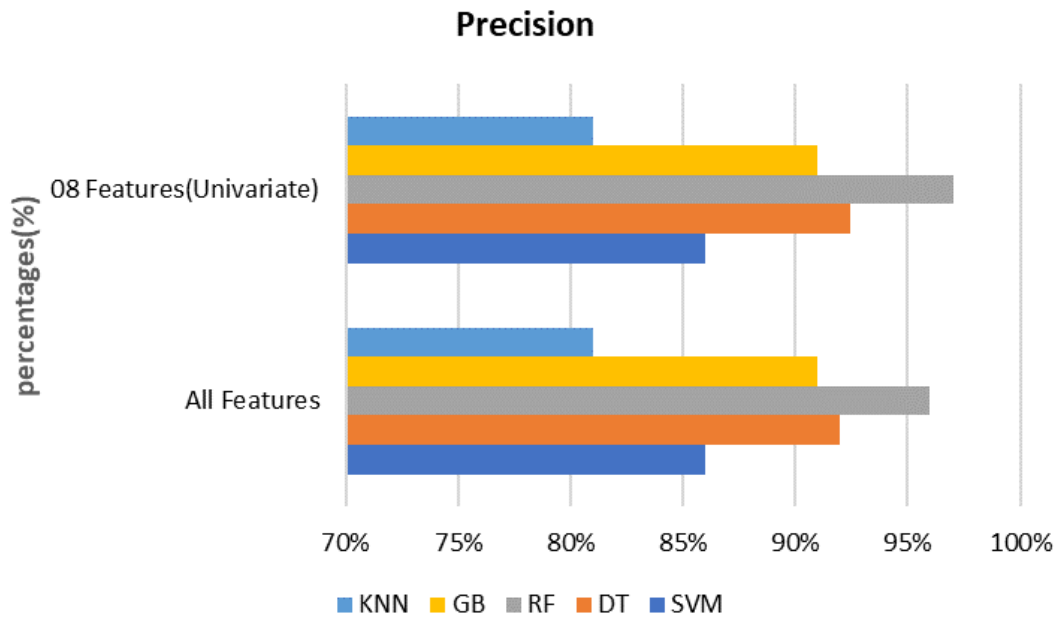


Figure 6.2: Precision For all 14(features) and 08 features

C. A COMPARISON BASED ON RECALL OF DIFFERENT METHODS:

A recall is obtained by apportionment the number of true positives by the total number of true positives plus false negatives. RF received a recall score of 96%, while KNN received a recall score of 81%. Similarly, assuming 14 input features, DT, GB, and SVM gained recall scores of 92%, 91%, and 86%, accordingly. In comparison to other classifiers, Random Forest obtains a higher recall score (98%) and KNN generates a relatively low recall score of 81%. Subsequently, the Univariate feature's recall for DT, GB, and SVM was 92%, 91%, and 86% accordingly.

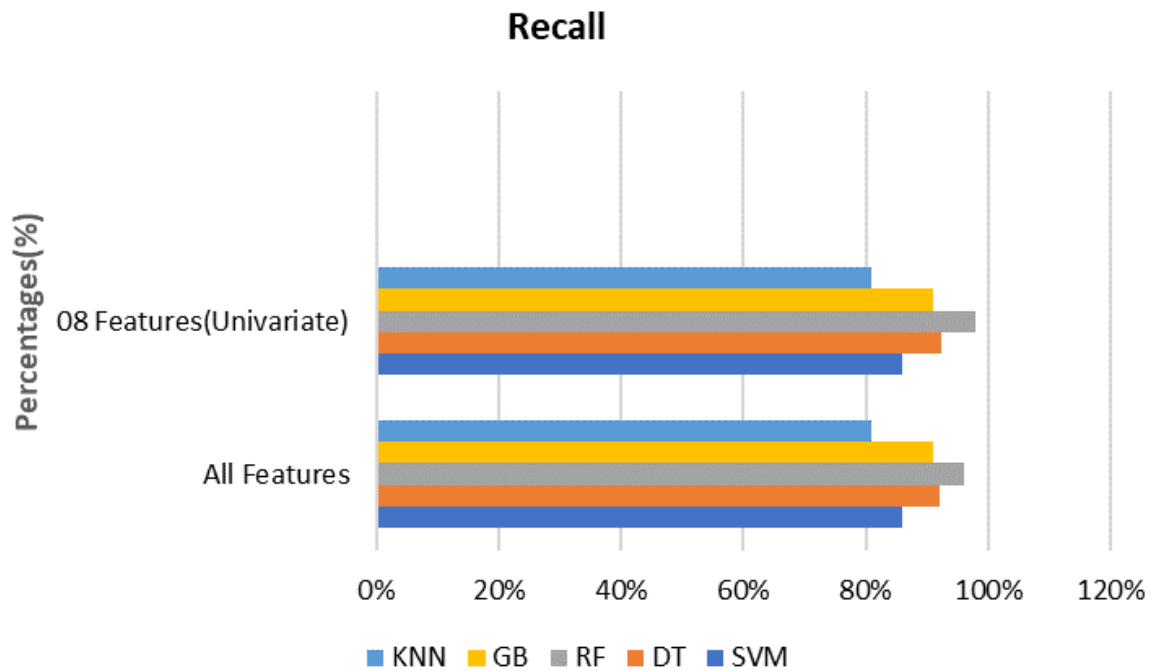


Figure 6.3: Recall For all 14(features) and 08 features

D. A COMPARISON BASED ON F1 SCORE OF DIFFERENT METHODS:

The relationship between Precision and Recall determines the F1 Score. While attempting to balance Precision and Recall, an F1 score is required. The maximum F1 score achieved by RF for the 14 features was 96%, and the weakest F1 score received by KNN was 81%. The F1 Scores of 92% and 91% are consistently acquired for DT and GB. SVM classifiers achieve an F1 score of 86%. KNN obtains a fairly poor F1 score of 81% when considering the five features that were chosen, while RF achieves the greatest F1 score of 96%. DT and GB scored 92% and 91%, respectively. Support Vector Machine scored an F1 score of 86%.

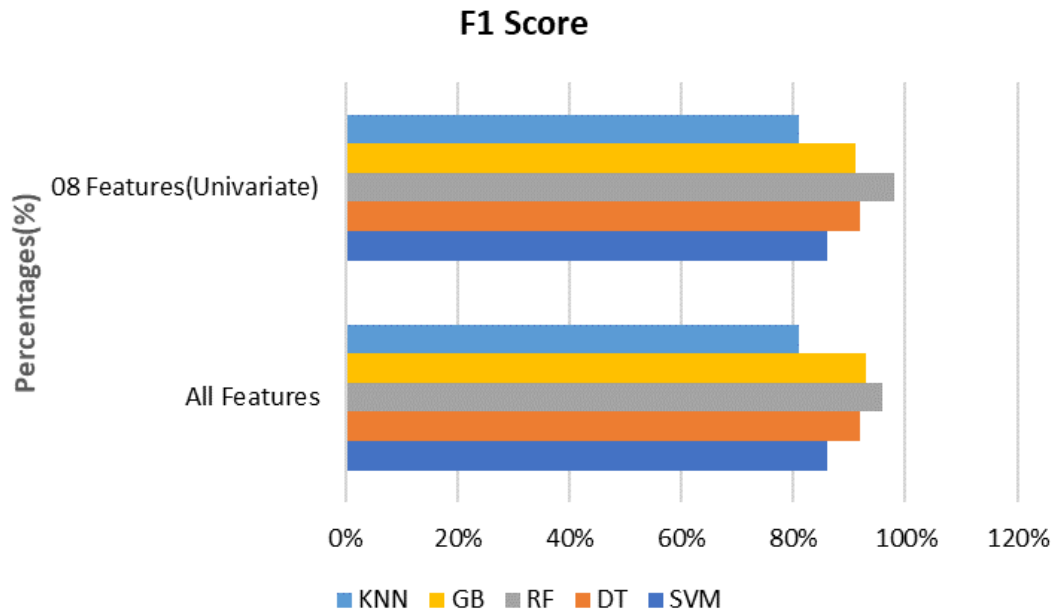


Figure 6.4: F1 Score For all 14(features) and 08 features

TABLE 5: PERFORMANCE COMPARISON OF FIVE CLASSIFIERS

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	98.31%	97.5%	98%	97.5%
Decision Tree	92.43%	92%	92%	91.5%
Gradient Boost	91.03%	90.5%	91%	90.5%
Support Vector Machine	86.27%	85.5%	86%	85.5%
K-nearest Neighbor	81.79%	81%	81.5%	81.5%

E. A COMPARISON BASED ON NEGATIVE PREDICTIVE VALUE OF DIFFERENT METHODS:

The percentage of people among those who receive a negative diagnostic test result who do not have the disease is known as the negative predictive value, which is a test performance feature. DT and RF classifiers were able to achieve the highest NPV (92.25%) in 14 features. Both K-nearest and gradient boosting achieve 90.14% NPV. The lowest NPV (85.21%) is obtained with SVM. With chosen Feature Univariate, DT,

RF, and GB scored the highest (92.25%) NPV. Both SVM and KNN receive (90.14%) NPV.

F. A COMPARISON BASED ON FALSE POSITIVE RATE OF DIFFERENT METHODS:

A subset of machine learning models allows for the measurement of an accuracy parameter known as the False Positive Rate. FPR is based on the number of real negatives that were predicted wrongly. Following the use of Univariate Feature Selection, SVM, GB, and KNN obtained 12.32% FPR. The false positive rate for DT and RF was 9.65%. Without using Univariate Feature Selection, SVM had the highest FPR (18.79%), and DT and RF had relatively low FPRs (9.65%). Additionally, both GB and KNN had an FPR of 12.32%.

G. A COMPARISON BASED ON FALSE NEGATIVE VALUE OF DIFFERENT METHODS:

An outcome that the model mistakenly forecasts belongs to the negative samples is referred to as a false negative value. The False Negative Rate for SVM, GB, and KNN with Univariate Feature Selection is 6.63%. RF and GB attained 5.18%. The maximum false negative rates for SVM without the feature selection technique are (10.09%). Both GB and KNN received 6.63% FNR. DT and RF were able to achieve a 5.18% False Negative Rate.

H. A COMPARISON BASED ON SENSITIVITY OF DIFFERENT METHODS:

The true positive rate, also known as sensitivity or recall, is a metric used in machine learning to determine the proportion of real positives that are accurately detected. Before using the univariate feature selection method, DT and RF had the greatest TPR (94.81%). GB and KNN achieved a True Positive Rate of 93.36%. The lowest TPR (89.90%) that SVM could attain. DT and RF had the same amount of TPR (94.81%) after utilizing the univariate feature selection technique. The True Positive Rate for SVM, GB, and KNN were all 93.36%.

I. A COMPARISON BASED ON SPECIFICITY OF DIFFERENT METHODS

The probability that the model would forecast negatively when the true value is negatively known as the true negative rate (Specificity, Selectivity). In our model, SVM has the lowest TNR of 81.20%, before using the univariate feature selection technique. GB and KNN obtain 87.67% of TNR. The highest number of True Negative Rates was obtained by DT and RF of 90.34%. Still 90.34% of TNR has Decision Tree and Random Forest after applying Univariate feature selection technique. SVM, GB, and KNN all obtained the same TNR, which is 87.67%.

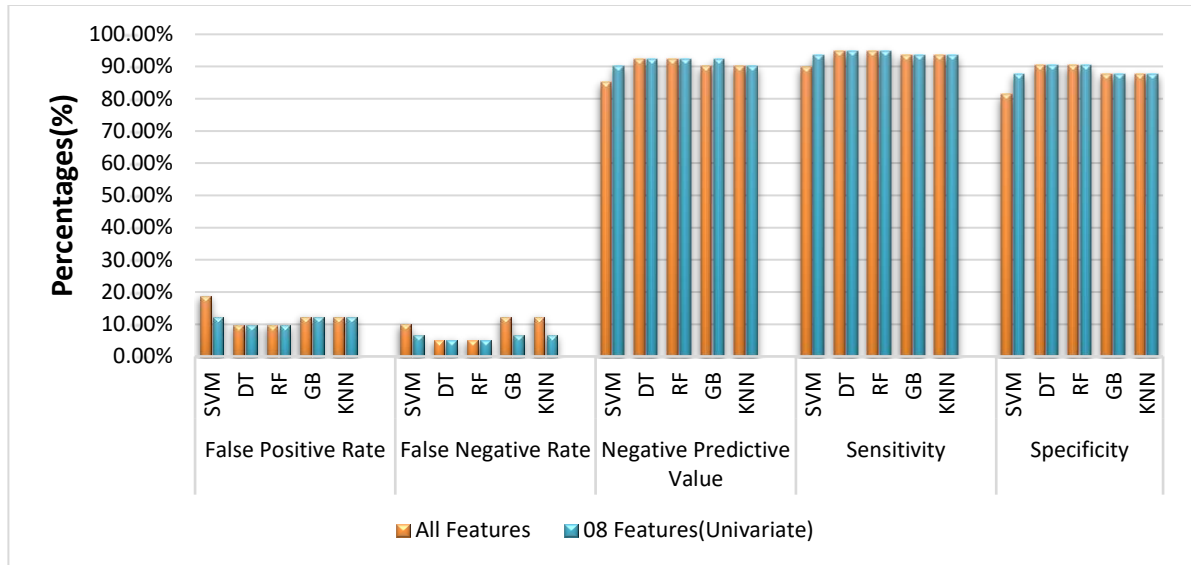


Figure 7: Results of FPR, FNR, NPV, Sensitivity & Specificity

4.3 Discussion

In this thesis paper, we present an analysis of using machine learning techniques to predict heart disease. Five different classifiers were evaluated: Decision tree, Gradient Boosting, N-Nearest Neighbors, Support Vector Machine, and Random Forest. The results of the study found that random forest provided the highest accuracy at 98.31%. We provided background information on heart disease and its significance as a major public health issue. They then detail the methods used in the study, including data pre-processing, feature selection, and model evaluation. The data used in the study was collected from a publicly available dataset, and various statistical methods were used to clean and prepare the data for analysis. Then we go on to present the results of the study, comparing the performance of each of the five classifiers. It was found that random forest had the highest accuracy among all the classifiers, with a 98.31% accuracy rate. We also discuss the reasons why random forest performed better than the other classifiers, highlighting its ability to handle a large number of features and its ability to handle both linear and non-linear relationships. Finally, we conclude by discussing the implications of the study and suggesting areas for future research. They indicate that machine learning techniques, particularly random forest, can potentially be used for heart disease prediction and could be used in clinical settings to improve the early detection and management of heart disease. In summary, this thesis paper thoroughly analyzes using machine learning techniques to predict heart disease. The results of the study indicate that Random Forest is the most effective classifier, with an accuracy of 98.31%. We provide valuable insights into the potential of machine learning in heart disease prediction and suggest areas for future research.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Early detection and prevention of heart disease can significantly improve patient outcomes because it is a leading cause of mortality and disability worldwide. The accuracy and effectiveness of cardiac disease prediction could be considerably increased by using machine learning techniques, which would have a significant influence on society. Potentially lessening the financial burden of heart disease on healthcare systems is one significant effect. Preventive actions can be made to lower the risk and delay the beginning of the disease by identifying those who are at a high risk of getting heart disease. Given that treating cardiac disease is frequently expensive and resource-intensive, this may result in cost savings for both people and healthcare systems.

Additionally, better patient outcomes may result from improved cardiac disease prognosis. Healthcare professionals can undertake early measures to prevent or lessen the burden of the condition by quickly identifying those who are at high risk of developing heart disease. Patients' quality of life may increase as a result, and the frequency of early heart-related fatalities may decline. Finding previously unknown risk variables for the disease is another potential effect of machine learning-based heart disease prediction. Machine learning algorithms may be able to find previously undetected patterns and associations by evaluating massive volumes of data. This could help us understand the root causes of heart disease better and help us design novel treatments and preventative measures.

Overall, applying machine learning to forecast cardiac disease has the chance to significantly advance illness detection, diagnosis, and care, with profound societal repercussions.

5.2 Impact on Environment

It is unclear how using machine learning to forecast cardiac disease would affect the environment specifically. However, the potential health advantages of enhanced cardiac disease prediction may have a variety of indirect effects on the environment. For instance, a decrease in the number of people who are unable to work or engage in physical activity owing to heart disease could result from its prevention and treatment. As individuals have the ability to be more active and rely less on driving, this may result in a decrease in the usage of fossil fuels for transportation. Additionally, as people can live and work for extended periods of time, there may be a decrease in the entire consumption of resources as a result of the increased longevity and health of those with heart disease.

On the other side, it's likely that using machine learning to forecast heart disease could have some unfavorable effects on the environment. For instance, the creation of computer

hardware can have a negative influence on the environment, as can the development and implementation of machine learning systems, which may necessitate the usage of resources like power and water. Moreover, the application of machine learning might necessitate the gathering and processing of massive volumes of data, which might necessitate the usage of server infrastructure that consumes a lot of energy.

Overall, even though it is unclear how machine learning for heart disease prediction may affect the environment directly, it is possible that the technology's potential health advantages will have either beneficial or negative indirect effects on the environment.

5.3 Ethical Aspects

When utilizing machine learning to forecast cardiac disease, there are a number of ethical issues to take into account. The possibility of bias in the creation and application of machine learning models is a crucial ethical consideration. Data is utilized to train machine learning models, and if the data is biased, the model may likewise be prejudiced. For instance, if a heart disease prediction model was trained primarily on data from one racial or ethnic group, the model may be less accurate for people belonging to other groups. According to the model's projections, people from underrepresented groups may be less likely to obtain the right preventative and therapeutic interventions, which could result in unequal access to care. Therefore, it is crucial from an ethical standpoint to ensure the variety and predictive ability of the data used to train machine learning models. Potential misuse of machine learning for prediction of heart disease is another ethical issue to take into account. For instance, insurers or employers might use the outcomes of heart disease prediction models to refuse coverage or employment to people who are thought to be at a high risk of contracting the illness. It is crucial to ensure that the outcomes of machine learning-based heart disease prediction are used ethically and in a way that respects people's rights and autonomy.

Moreover, the use of machine learning to the prediction of heart disease raises concerns about the importance of technology in healthcare and the possibility of its eventual replacement of human healthcare workers. It is crucial to take into account the potential effects on healthcare professionals and to make sure that machine learning is utilized to complement human expertise rather than to replace it.

The potential for bias and abuse, the use of technology in healthcare, and the need to respect people's rights and autonomy are the main ethical issues surrounding the use of machine learning to predict heart disease.

5.4 Sustainability Plan

The procedures that will be done to guarantee that the use of the method is sustainable over the long run are outlined in a sustainability strategy for the use of machine learning for heart disease prediction. This can entail taking steps to address the technology's effects on the environment, society, and the economy.

Measures to reduce the environmental impact of using machine learning to forecast heart disease should be part of a sustainability plan. This could involve making steps to utilize less resources, such water, and less energy, such as employing gear and servers that are energy efficient. In regard to social longevity, a strategy might address concerns about the likelihood of bias and technology abuse as well as possible effects on healthcare professionals. This could involve making ensuring that the dataset used to train machine learning models is diverse and representative, as well as making sure that the models' outputs are used morally and in a way that respects people's autonomy and rights.

A sustainability strategy should include efforts to guarantee that the technology is affordable and beneficial to all stakeholders in order to ensure the economic sustainability of using machine learning to forecast cardiac disease. This may entail making efforts to cut expenses connected with the creation and application of the technology as well as ensuring that the advantages of the technology are equitably distributed among all stakeholders.

In general, a sustainability strategy for using machine learning to forecast heart disease should address the technology's effects on the environment, society, and the economy and lay out the activities that will be made to assure its long-term use.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION & IMPLICATION FOR FUTURE WORK.

6.1 Summary of the Study

Finding the finest machine learning algorithm that can effectively predict heart illness is the main goal of this system. The efficiency of classifiers employing both the original features and the features selected using a univariate feature selection technique is investigated. The dataset is divided into training and testing after the feature selection technique is applied. The training phase will use 70% of the data, and the testing phase will use the remaining 30%, in accordance with model learning rates. All ensemble methods with classifications are created in order to compare results across the combined dataset. The chosen features are subjected to the application of various machine learning classification approaches, including SVM, DT, RF, GB, and KNN. RF had an accuracy of 98.31% with 14 characteristics.

6.2 Conclusion

We have experimented in this research project to determine how well different ML algorithms perform. More conventional regression procedures can be improved with the use of these algorithms in terms of diagnosis and prognosis. Cardiovascular disease prediction is a serious issue for people, hence one of the criteria used to evaluate an algorithm's effectiveness is its accuracy. The diagnosis of cardiac-related disorders can be greatly impacted by a little improvement in prediction accuracy. Using Univariate feature selection as the basis for feature selection strategies, we examine the dataset's characteristics in this article and select the optimum collection of features. The chosen features are subjected to the application of a number of Classification using machine learning techniques, including SVM, DT, RF, GB, and KNN. Several machine learning classification techniques, such as SVM, DT, RF, GB, and KNN are applied to the selected features. With 14 features, RF had an accuracy of 98.31%. To reduce the rate of mortality cases through increased disease awareness, more machine learning techniques will be deployed in the future to analyze cardiac problems more effectively and detect illnesses early.

6.3 Implication for Further Work

Following are a few implications that could emerge from additional research utilizing this heart disease:

1. This thesis can be applied to the creation of mobile applications or websites for users.
2. To support this thesis, additional disease kinds can be included.
3. To enhance this thesis's results, other datasets may be incorporated.
4. Different types of classifiers can be implemented.

REFERENCES:

- [1] Ahmed, H., M.G. younis, E., Hendawi, A., & Abdelmgeid A. Ali, A. (2020, October). Heart disease identification from patients' social posts, machine learning solution on Spark. *Future Generation Computer Systems*, 111,714-722. doi:<https://doi.org/10.1016/j.future.2019.09.056>
- [2] Stamate D, Alghamdi W, Ogg J, Hoile R, Murtagh F. (2018). A Machine Learning Framework for Predicting Dementia and Mild Cognitive Impairment. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 671-678).
- [3] Chaurasia V, Pal S, Tiwari BB.(2018). Chronic kidney disease: a predictive model using decision tree. *International Journal of Engineering Research and Technology*.
- [4] P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques," in *IEEE Access*, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
- [5] F. Z. Abdeldjouad, M. Brahami, and N. Matta, *A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques*. Cham, Switzerland: Springer, 2020, pp. 299–306.
- [6] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2017, pp. 204–207.
- [7] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics Informat.*, vol. 36, pp. 82–93, Mar. 2019.
- [8] M. Ashraf, S. M. Ahmad, N. A. Ganai, R. A. Shah, M. Zaman, S. A. Khan, and A. A. Shah, *Prediction of Cardiovascular Disease Through Cutting-Edge Deep Learning Technologies: An Empirical Study Based on TENSORFLOW, PYTORCH and KERAS*. Singapore: Springer, 2021, pp. 239–2
- [9] D. Singh and J. S. Samagh, "A comprehensive review of heart disease prediction using machine learning," *J. Crit. Rev.*, vol. 7, no. 12, p. 2020, 2020.
- [10] Aggrawal, R., & Pal, S. (2021). Elimination and Backward Selection of Features (P-Value Technique) In Prediction of Heart Disease by Using Machine Learning Algorithms. *Turkish Journal of Computer and Mathematics Education*, 12, 2650-2665.
- [11] Javid, A. Khalaf, and R. Ghazali, "Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method," *Int.J.Adv. Comput. Sci.Appl.*, vol.11, no.3, 2020.
- [12] D. Singh and J. S. Samagh, "A comprehensive review of heart disease prediction using machine learning," *J. Crit. Rev.*, vol. 7, no. 12, p. 2020, 2020.
- [13] C. Trevisan, G. Sergi, S. J. B. Maggi, and H. Dynamics, "Gender differences in brain-heart connection," in *Brain and Heart Dynamics*. Cham, Switzerland: Springer, 2020, p. 937.
- [14] M. S. Oh and M. H. Jeong, "Sex differences in cardiovascular disease risk factors among Korean adults," *Korean J. Med.*, vol. 95, no. 4, pp. 266–275, Aug. 2020.

- [15] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *Int. J. Pharmaceutical Res.*, vol. 12, no. 4, 2020.
- [16] R. Bhuvaneeswari, P. Sudhakar, and G. Prabakaran, "Heart disease prediction model based on gradient boosting tree (GBT) classification algorithm," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2, pp. 41–51, Sep. 2019.
- [17] J. Cheng, G. Li, and X. Chen, "Research on travel time prediction model of freeway based on gradient boosting decision tree," *IEEE Access*, vol. 7, pp. 7466–7480, 2019, doi: 10.1109/ACCESS.2018.2886549.
- [18] A. U. HAQ et al., "Identifying the Predictive Capability of Machine Learning Classifiers for Designing Heart Disease Detection System," 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, 2019, pp. 130-138, doi: 10.1109/ICCWAMTIP47768.2019.9067519.
- [19] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
- [20] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
- [21] G. O.Rashmi and U. M.A. kumar, "Machine learning methods for heart disease prediction," *Int.J.Eng.Adv. Technol.*, vol.8,no.5S,pp.220–223, May 2019.
- [22] Beunza, Juan-Jose, et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)." *Journal of biomedical informatics* 97 (2019): 103257.

APPENDIX RESEARCH REFLECTION

An interesting and difficult project that called for a variety of data science skills and knowledge was the process of utilizing machine learning to predict heart disease. Getting a big, diverse dataset of people with known heart disease status to train and test the model on was one of the main hurdles. In order to ensure that the data was prepared for analysis, this needed intensive data collecting and preprocessing, including cleaning and normalizing the data. Finding the most pertinent and predictive variables to incorporate into the model after the data had been prepared was the next difficult task. The input data had to be carefully chosen and prepped, and new features had to be created using feature engineering approaches. Overall, working on this project gave me the opportunity to gain a wealth of information and abilities in the area of ML. It offered a chance to put these skills to use and advance them while also learning about the difficulties and factors related to applying machine learning to forecast cardiac disease.

EHDP5C

ORIGINALITY REPORT

25% SIMILARITY INDEX	20% INTERNET SOURCES	19% PUBLICATIONS	13% STUDENT PAPERS
--------------------------------	--------------------------------	----------------------------	------------------------------

PRIMARY SOURCES

1	ris.cdu.edu.au Internet Source	2%
2	Submitted to Daffodil International University Student Paper	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
4	turcomat.org Internet Source	2%
5	www.philstat.org.ph Internet Source	1%
6	www.researchgate.net Internet Source	1%
7	Submitted to University of Technology, Sydney Student Paper	1%
8	Pronab Ghosh, Sami Azam, Mirjam Jonkman, Asif Karim et al. "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO	1%