CARDIOVASCULAR DISEASES (CVD) PREDICTION MODELS: A SYSTEMATIC REVIEW

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

Bioinformatics

By

Raghav Luthra (191903)

Under the supervision of

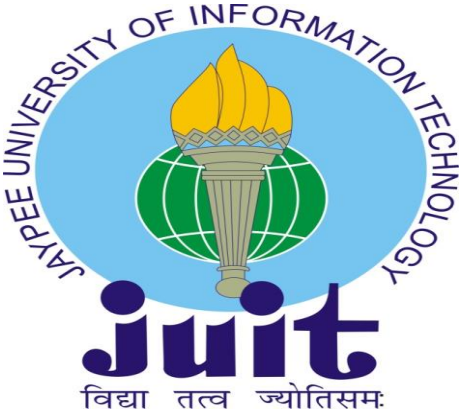
Dr. Gopal Singh Bisht

and

co-supervision of

Dr. Vipul Kumar Sharma

to



Department of Biotechnology & Bioinformatics

Jaypee University of Information Technology, Waknaghat, Solan-173234, Himachal Pradesh

Certificate

Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Cardiovascular diseases (CVD) prediction models: a systematic review"** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Bioinformatics** submitted in the Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from August 2022 to May 2023 under the supervision of **Dr. Gopal Singh Bisht** (Associate Professor, Department of Biotechnology and Bioinformatics) and co-supervision of **Dr. Vipul Kumar Sharma** (Assistant Professor (SG), Department of Computer Science Engineering & Information Technology).

I also authenticate that I have carried out the above-mentioned project work under the proficiency stream **Data Science**.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Raghav Luthra, 191903

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Gopal Singh Bisht

Associate Professor

Department of Biotechnology & Bioinformatics

Dated: 8 May 2023

Dr. Vipul Kumar Sharma

Assistant Professor (SG)

Department of Computer Science & Information Technology

Dated: 8 May 2023

Plagiarism Certificate

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: .....

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: _____ **Department:** _____ **Enrolment No** _____

Contact No. _____ **E-mail.** _____

Name of the Supervisor: _____

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): _____

_____

_____

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at .....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

<table>
<tr><td>Copy Received on</td><td rowspan="4">• All Preliminary Pages
• Bibliography/Images/Quotes
• 14 Words String</td><td rowspan="3"></td><td>Word Counts</td><td></td></tr>
<tr><td>Report Generated on</td><td>Character Counts</td><td></td></tr>
<tr><td rowspan="2"></td><td>Submission ID</td><td>Total Pages Scanned</td><td></td></tr>
<tr><td></td><td>File Size</td><td></td></tr>
</table>

Checked by
Name & Signature **Librarian**

.....

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

Acknowledgement

I will start this report by acknowledging my deepest gratitude for the divine spiritual aura that takes selfless care of me and has bestowed me with a lot of great opportunities in life. I will also take this opportunity to thank my parents, especially my father who has supported me tirelessly during the entire tenure of my Bachelors. Next, I would like to mention my indebtedness and gratitude to my project supervisor, Dr. Gopal Singh Bisht, Associate Professor, Department of BT & BI and my co-Supervisor Dr. Vipul Kumar Sharma, Assistant Professor (SG), Department of CSE & IT. Their endless patience, constant scholarly guidance, continual encouragement along with their constructive criticism especially while helping me draft my report at all stages has only made it possible for me to complete this project. I would like to express my heartiest gratitude to both, Dr. Gopal and Dr. Vipul, for guiding me and laying the technical ground work; their constant guidance is the reason I could attempt and finish this inter-disciplinary project. I will also take this opportunity to thank my friends, administration, staff and the lab access our university has provided for their convenient help. Thank you.

Table of Contents

List of Figures

List of Tables

Abstract

Cardiovascular diseases (CVD) have the highest mortality rate in the Indian healthcare system. The aim of this study and report is to collect medical data related to cardiovascular diseases and extract the maximum features and implement them in machine learning based algorithms to predict CVD risk. This systematic review is rich in data visualisation and model implementation along with an exhaustive analysis of performance metrics especially sensitivity and specificity analysis such as accuracy, precision and recall. The usefulness and utility of the best model that can accurately capture data and effectively predict the risk of CVD is worked on in this report.

CHAPTER 1: INTRODUCTION

1.1 Introduction

Heart is the most vital organ of the human body. It supplies blood to every organ and bone in our body, which is the base line of a human's survival. If the heart, dysfunctions or has anomaly in any way, the brain alongside other organs will be drastically affected and eventually stop functioning, and the individual will die in a matter of minutes. An array of heart-related disorders, specifically termed Cardiovascular Diseases (CVDs) are becoming more common due to changes in lifestyle, increased personal and professional stress, or in other words poor stress management in the fast-paced lifestyle these days along with depleted nutrition and bad eating habits [1]. Therefore, it is the need of the hour to analyze this alarming issue, hence making it imperative that CVDs be accurately and practically predicted using data visualization, analysis and subsequent classification techniques.

Early detection and accurate prediction of CVDs are crucial for effective prevention and management. Machine learning (ML) algorithms have shown great potential in predicting CVDs, based on clinical data, medical imaging, and genetic information [2]. Machine learning (ML) techniques have the potential to improve the accuracy of CVDs prediction, enabling earlier detection and more effective treatment.

Machine learning is a subfield of artificial intelligence (AI) that involves using statistical models and algorithms to enable computers to learn from data without being explicitly programmed. ML techniques can be used to analyse large volumes of data to identify patterns and predict outcomes [3]. In the case of CVDs prediction, ML models can be trained on a variety of patient data, including demographic information, medical history, and clinical and laboratory test results, to generate predictive models that can identify individuals who are at high risk of developing CVDs.

Worldwide, medical organizations keep a tab on patient history and medical information, eventually compiling useful technical information on a range of various health-related topics, under the branch of research and development. Some datasets are restricted/private while some have open access i.e. are open source. The amount of medical imaging and statistical data gathered is typically enormous, and this data is

quite noisy. For maximum feature extraction, these datasets must be gathered effectively and processed in a systematic way.

Once the data dispersion with respect to the relevant parameters is understood, collected datasets are trained using the feasible features. A predictability layer is, then added, which can in turn be effectively integrated to design a predictive classifier. After analyzing the data, it is necessary to study and implement supervised and unsupervised classification paradigms to primarily understand data dispersion and examine the potential these datasets exhibit, which are too large for human minds to process. However, the quality and completeness of these datasets may vary, and pre-processing is necessary to ensure the accuracy and consistency of the data.

The pre-processing steps typically include data cleaning, feature selection, and normalization. Data cleaning involves identifying and removing missing, inconsistent, or erroneous data points [4]. Feature selection aims to identify the most informative variables that are relevant to the prediction task and reduce the dimensionality of the dataset [5]. This is important because high-dimensional data can lead to overfitting and reduced generalization performance. Normalization is used to scale the variables to a common range, which helps to improve the performance of ML algorithms and reduce the impact of outliers [6].

There are several different types of ML algorithms that can be used for CVDs prediction, including supervised learning, unsupervised learning, and reinforcement learning [7]. Supervised learning algorithms involve training models on labelled data, where the outcome of interest (e.g., whether or not a patient has CVDs) is known. Unsupervised learning algorithms are used when there is no labelled data available, and the algorithm must identify patterns and relationships in the data on its own. Reinforcement learning involves training models to make decisions based on feedback from the environment.

One of the most commonly used ML techniques for CVDs prediction is logistic regression. Logistic regression is a type of supervised learning algorithm that can be used to predict the probability of an event occurring (e.g., developing CVDs) . Logistic regression models are trained on a set of independent variables (e.g., age, sex, blood pressure, cholesterol levels) and a dependent variable (e.g., whether or not a patient has

CVDs). The model then generates a predicted probability of CVDs based on the input variables [8].

Another popular ML algorithm for CVDs prediction is decision trees. Decision trees are a type of supervised learning algorithm that uses a tree-like structure to make decisions based on input variables. Each node in the tree represents a decision based on a specific variable, and the leaves of the tree represent the predicted outcomes [9]. Decision trees are often used in combination with other ML techniques, such as random forests, to improve their accuracy.

Artificial neural networks (ANNs) are another ML technique that has been used for CVDs prediction. ANNs are a type of supervised learning algorithm that is modelled on the structure of the human brain. ANNs consist of layers of interconnected nodes, each of which performs a simple calculation. The output of one layer is fed as input to the next layer, and the network learns to adjust the weights of the connections between the nodes to improve its accuracy [10]. ANNs have shown promising results in predicting CVDs, particularly when combined with other ML techniques.

Reinforcement learning can further be applied to refine the accuracy of prediction models, by taking live feedback. Medical diagnosis is a crucial yet challenging art form that must be performed precisely and effectively.

The performance of ML models in CVD prediction is typically evaluated using various metrics, such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC), and F1 score. Accuracy measures the proportion of correctly predicted instances, while sensitivity measures the proportion of true positive instances that are correctly identified by the model [11]. Specificity measures the proportion of true negative instances that are correctly identified by the model. AUC-ROC is a metric that summarizes the trade-off between sensitivity and specificity across different thresholds of the predicted probabilities. F1 score is a harmonic mean of precision and recall, which measures the balance between the positive and negative predictive values [12].

One of the main challenges in using ML techniques for CVDs prediction is the availability and quality of data. ML algorithms require large volumes of high-quality data to be trained effectively. Additionally, data must be collected from diverse

populations to ensure that the predictive models are accurate across different demographic groups. Another challenge is the interpretability of ML models.

Post a systematic survey, research can be continued in the direction of automating the entire system which will be extremely contributing in cardiovascular disease prediction. This study is an inter disciplinary collaboration between biotechnology, biosciences & computer science and information technology systems.

1.2 Problem Statement

A traditional reporting method for diagnosing CVD is based on the patient's medical history, physical examination by a physician and analyzing subsequent results based on invasive and noninvasive lab tests. These traditional methods are excessively time consuming and can even result in inaccurate diagnosis, if the history is wrong or there are any discrepancies in physical/lab examinations. In addition, these methods are very costly and in fact computationally intensive to diagnose the disease(s). In a country like India, where there is a huge population, not every person has enough awareness, time, resources or money for the diagnosis, let alone following up with the required treatment. An attempt to study and implement data encapsulation, visualization along with feature extraction based on CVD datasets and **exploiting machine learning algorithms for classification and accuracy assessment** has been ventured using a Kaggle dataset and classification implemented in Python.

1.3 Objectives

- Dataset review: Collecting and understanding CVD based dataset along with its pre-processing, data cleaning and data visualisation w.r.t an array of CVD parameters (features) for cardiovascular disease prediction.
- Study and implementation of machine learning based classifiers to accurately detect CVDs.
- Comprehensive accuracy assessment summary and comparison between the classifiers to provide an overview of prediction models for the risk of CVD.

1.4 Methodology

The methodology plan comprises of data acquisition (wise choice of dataset), dataset preparation (data cleaning, missing values, data integration, data reduction, pre-processing and labelling), choosing proper machine learning based approach for CVD and then training them by using both supervised and unsupervised learning. Predictive models such as K-nearest neighbour (KNN), Support Vector Machines (SVM), Gaussian Naïve Bayes, Decision trees, and Random forests along with logistic regression are implemented on CVD dataset which has features such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, ECG results at rest, maximum heart rate during stress test, exercise induced angina, old peak, ST slope and target.

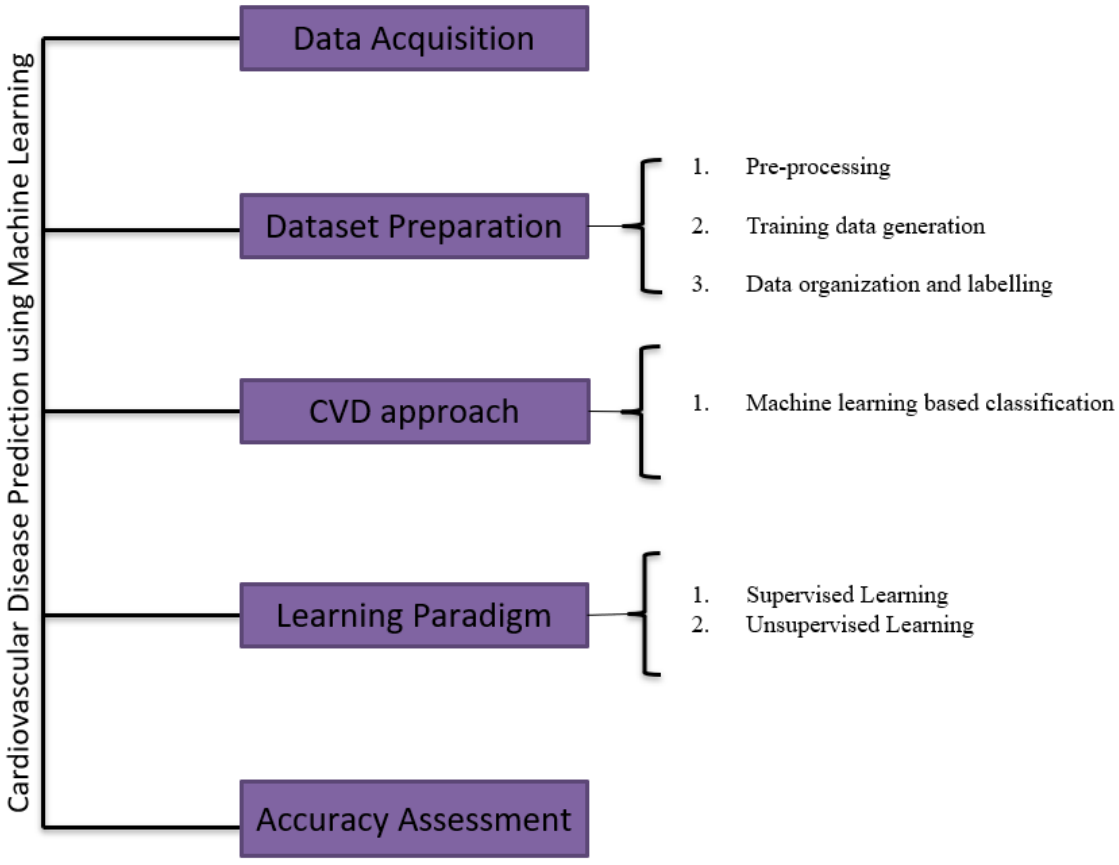


Figure 1: Cardiovascular diseases prediction using Machine Learning

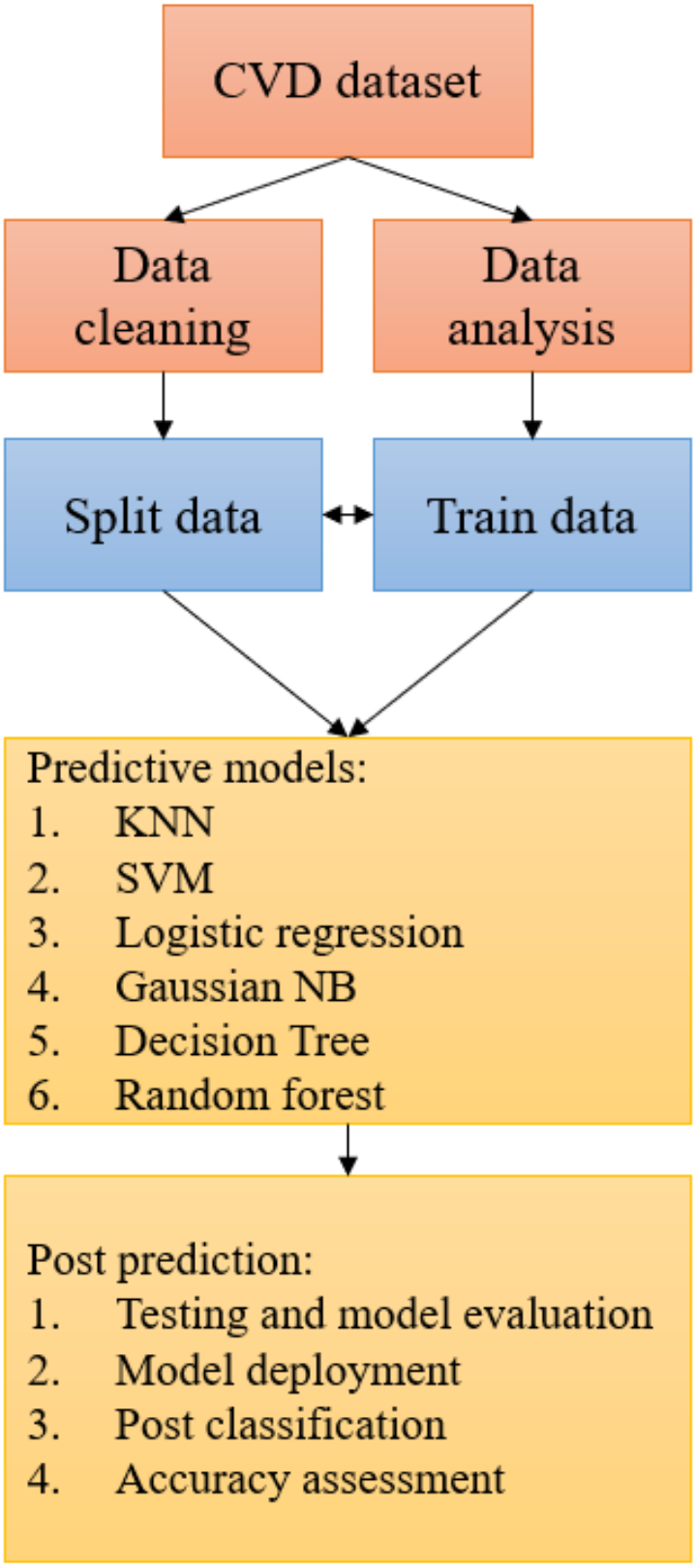


Figure 2: Proposed methodology for CVD prediction

1.5 Organization

The report has been drafted and organised carefully keeping in mind the research objectives. Chapter 1 of the report deals with the introduction to cardiovascular diseases (CVD) and gives an insight into the traditional monitoring and diagnosis scheme, while defining the problem statement and the objectives, along with a glimpse of the data set and methodology implemented. Chapter 2 is an in-depth literature review of the current state of the art revolving around Cardiovascular diseases (CVD) prediction models. Chapter 3 expands the materials and methods used to implement our research objectives while the results are discussed in Chapter 4 of the report.

CHAPTER 2: LITERATURE SURVEY

A detailed state of the art and thorough literature review was performed on the topic to understand the problem statement and its associated history, research trends and chronological upgradation of the techniques regarding CVDs. There seems to be no doubt that CVD is one of the prime reasons of mortality around the globe [13]. Heart disease cases are rising quickly every day, thus it's crucial and worrisome to predict any potential illnesses in advance. This diagnosis is a challenging task that requires accuracy and efficiency [14]. Most nations lack cardiovascular specialists, and a high percentage of cases are misdiagnosed. These issues could be resolved by creating a precise and effective early-stage heart disease prediction tool that uses digitised patient records to enhance clinical decision-making [15]. The healthcare structure is very rich information and data wise but the knowledge extraction is still a work in progress, this is majorly due to lack of effective tools in the analysis and visualisation department, as well as the inability to map hidden relationships and trends in CVD data [16]. Reliable and feasible diagnosis system aided with machine learning based classifiers can help curate quicker prediction [17]. Automation of models trained and tested on real time data with varied target attributes for better prediction [16] [18]. When it comes to machine learning based predictive algorithms, it has been observed that they have been very effective in learning inter-data patterns and relationships [19].

Decision trees in a certain dataset performed really well when they were used with Principal Component Analysis (PCA), but gave very poor results in some other cases which could have been due to overfitting [18]. Random Forest Classifier performed extremely well because it could solve the problem of overfitting by using multiple algorithms i.e. multiple decision trees [18]. SVM and Naive Bayes classifier were computationally very fast and performed really well in most of the cases [18]. In another dataset To reduce the large volume of data, also known as big data, data mining can be used to reduce the complexity of the data [20]. A recent study aims to give a thorough explanation of how Naive Bayes and decision tree classifiers are used in our research, especially when it comes to the prediction of heart disease. The results of an experiment comparing the use of predictive data mining techniques on the same dataset show that Decision Trees perform better than Bayesian classification [21]. In order to increase the accuracy of heart disease classification, another researcher used the Fast Correlation-Based Feature Selection (FCBF) method to filter out duplicate information [22]. The optimisation hybrid technique increases the predicted accuracy of medical data sets, according to

experimental results [22]. Techniques such as support vector machines, KNNs, genetic algorithms, perceptron neural networks, decision tress (DT), random forests (RF) have been discussed exhaustively over the span of years [3] [6]. A survey report on the prediction of cardiac ailments has demonstrated that hybridization performs well and provides better prediction accuracy than the older machine learning techniques [24]. Different features of CVD are clubbed and trained in various classifier models as well [13], [25].

CHAPTER 3: MATERIALS AND METHODS

3.1 Material & Methods

Dataset acquisition and pre-processing: Collecting and understanding CVD based dataset along with its pre-processing, data cleaning and data visualisation w.r.t an array of CVD parameters (features) for cardiovascular disease prediction. The cardiovascular disease dataset studied in this report is acquired from Kaggle. In the dataset, there are twelve columns which are described in detail below [26].

1. **Age**: Individual's age
2. **Sex**: Gender of the individual:
   1 = Male
   0 = Female
3. **Chest-pain type (cp)**: Shows the type of chest pain that the individual experiences:
   0 = Typical angina
   1 = Atypical angina
   2 = Non anginal pain
   3 = Asymptotic
4. **Resting Blood Pressure (trestbps)**: Shows the resting blood pressure of the individual (mmHg). (Range = 90/60mmHg – 120/80mmHg).
5. **Serum Cholesterol (chol)**: Serum cholesterol (mg/dL). Optimal is less than 100mg/dL.
6. **Fasting Blood Sugar (fbs)**: Displays the fasting blood sugar of an individual.
   In case fasting blood sugar is greater than 120mg/dL then enter: 1 (true)
   else: 0 (false).
7. **Resting ECG (restecg)**: Shows resting electrocardiographic results
   0 = Probable left ventricular hypertrophy
   1 = Normal
   2 = Abnormalities in the T wave and ST segment.
8. **Maximum heart rate during stress test**: (220 – age of person) [27].
9. **Exercise induced angina (exang)**: Enter 0 if the person does not have angina during exercise otherwise enter 1 if the person has angina during exercise.

10. **Old Peak**: ST depression induced by exercise relative to rest. (Range = 0 – 0.62).
11. **ST Slope**: Enter 0 if the slope of ST segment is descending, enter 1 if the slope of ST segment is flat and enter 2 if the slope of the ST segment is ascending.
12. **Target**: Enter 1 if the person has heart disease and enter 0 if there is no heart disease.

Following is the screenshot of the dataset which shows all the columns and the number of rows which are present in the dataset which has been acquired from Kaggle:

	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1185	45	1	1	110	264	0	0	132	0	1.2	2	1
1186	68	1	4	144	193	1	0	141	0	3.4	2	1
1187	57	1	4	130	131	0	0	115	1	1.2	2	1
1188	57	0	2	130	236	0	2	174	0	0.0	2	1
1189	38	1	3	138	175	0	0	173	0	0.0	1	0

1190 rows × 12 columns

Figure 3: Screenshot of Dataset

Study and implementation of machine learning based classifiers to accurately detect CVDs:

The classifiers used to accurately detect CVDs in this study are mentioned below:

- Logistic Regression
- KNN
- Random Forest Classifier
- SVM
- Decision Tree
- Gaussian Naïve Bayes

As the project is implemented using python, noteworthy libraries that are used are:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Skicit-learn

Numpy is the fundamental scientific package of python which was created specifically for computing and processing one and multidimensional arrays [28]. Pandas can be used to explore the economic relationships among many types of data using forecasting as a foundation and is chosen over alternatives because it is simpler, faster, and more expressive than any of its rivals [29]. Matplotlib, on the other hand aids data visualisation and is Python's visualisation library for 2D displays of arrays, all our research plots of data visualisation has been implemented using this package [30]. Seaborn, which is based on Matplotlib only, is used for graphical and statistical visualisation [31]. SVM, random forests, gradient raises, k-means, and a good deal of more categorization, retransmission, and integration algorithms were created to operate with Python's NumPy and SciPy libraries, which are built to work with numerical and scientific data in a significant way and aided in our study.

3.2 Models studied and implemented

1. Logistic Regression

Used for classification and predictive analysis, this technique is used to forecast the results of the categorical attributes; for instance, the results in such cases may be either true or false, yes/no, 0 or 1, etc [32]. However, it does not provide an exact output; instead, it provides a value between 0 and 1 [33]. It can categorise datasets that are both continuous and discrete, making it a crucial ML algorithm.

Types of logistic regression:

- Binomial: As the name suggests, this one has only two outcomes in this case, yes or no, 0, 1, etc.
- Multinomial: It allows for the creation of unordered results of more than two different kinds, such as the names of cars, animals, etc.
- Ordinal: It allows for the achievement of more than three different kinds of ordered results, such as, "low," "medium," and "high."

2. KNN

KNN is a non-parametric supervised learning technique that is used for both classification and regression [34]. It is also very sensitive to the local structure of the data involved and it doesn't quickly pick up knowledge from the training set, instead, it makes a change to the dataset during categorization [35].

Functioning of KNN:

- Choose the number of cases in step 1 or the number of “k”
- Determine the Euclidean separation between K nearest neighbours
- Separate the cases based on the Euclidean distance that was determined
- In these circumstances, data points are partitioned
- Put the newly acquired data points in the group with the most neighbours

KNN is easy to implement and is very robust for noisy datasets though it can be computationally complex to work around [35].

3. Random forest classifier

A well-liked and widely implemented supervised learning method that fits a number of decision tress on a huge sub-samples is called a random forest [36].

It is a classifier that use numerous decision trees for various dataset subsets before averaging the outcomes to raise the predicted accuracy for the dataset, it also gathers predictions from each decision tree rather than depending solely on one, and forecasts the outcome based on the votes of the majority of the predictions [37].

A random forest classifier can be further improved by tuning the hyperparameters such as tree size, tree depth, min samples, etc.

4. SVM

Support vector machine, or SVM, is a type of supervised machine learning whose objective is to create the best possible partitions for classifying objects in n-dimensional space [38]. It can perform both linear as well as nonlinear classification [39].

Types of SVM:

1. Linear: A single line is used to divide the dataset into two sections. This kind of dataset is referred to as linearly separable data, and the linear SVM classifier is used.
2. Non-linear: A straight line is used to divide a dataset, in which case it is referred to as non-linear data, and classifiers applied to such datasets are referred to as non-linear classifiers.

5. Decision Tree classifier

Decision tree (DT) is a supervised machine learning which has the concept of nodes to determine the outputs, the first is a group of nodes called decision nodes, which are used to make decisions while the leaf nodes display the outcomes [40]. Decision tree is great for both numerical and category values and as it mimics the structure of a tree, hence the name, CART (Classification and Regression tree algorithm) is used to build a tree [41].

6. Gaussian Naive Bayes

A generative supervised model which supports continuous data, this categorization technique makes use of the Bayes theorem; here the Gaussian Naive Bayes assumes independence of the features or parameters and it has class-specific covariance matrices [42]. Naive Bayes is a classification method. It can handle complex, dependent, nonlinear data, making it suited for the heart disease dataset [43].

Comprehensive accuracy assessment summary and comparison between the classifiers to provide an overview of prediction models for the risk of CVD:

After pre-processing the data and running them through six machine learning based classifiers, accuracy assessment is performed. Performance metrics play a key role in determining the level of classification and distinctiveness a model has offered. The key performance metrics in this case being confusion matrix, precision and recall, which perform much better than just accuracy and give us deep performance insights [44]. They have been summarised in the equations below:

TPR = (TP / Actual positive) = TP / (TP + FN) ……..(1)
FNR = (FN / Actual Positive) = FN / (TP + FN) ……(2)
TNR = (TN / Actual Negative) = TN / (TN + FP) ……(3)
FPR = (FP / Actual Negative) = FP / (TN + FP) ……(4)
Precision = TP / (TP + FP) ……(5)
Recall = TP / (TP + FN) ……..…..(6)
Accuracy = (TP + TN) / (TP + TN + FP + FN) …….. (7)
F1-score = 2 * (Precision * Recall) / (Precision + Recall)……..…..(8)

Confusion Matrix: Also known as error matrix, it is one of the popular summative performance metrics to evaluate the overall performance of a machine learning based classifier [45]. Simply put, it is a comparative analysis in the form of a mathematical matrix which compares the ground truth values, i.e., actual values with respect to the classifier's predicted values. The four quadrants in a confusion matrix are as follows:

1. **TP (True Positive)** = Model correctly predicts the positive class, you predicted positive and its true.

2. **TN (True Negative)** = Model correctly predicts the negative class, you predicted negative and its true.
3. **FP (False Positive)** = Model gives the wrong prediction of the negative class, you predicted positive but its false.
4. **FN (False Negative)** = Model wrongly predicts the positive class, you predicted negative but it's true.

In machine learning, once the contingency table is computed several other metrics can be calculated and assessed for the classifier's performance. The following terms are used to evaluate the performance of a binary classification model:

1. **TPR (True Positive Rate)** = It is also known as sensitivity or recall and is the amount of actual positive instances that the model accurately defined as positive.
2. **FNR (False Negative Rate)** = It is defined as the amount of actual positive instances that the model incorrectly defined as negative.
3. **TNR (True Negative Rate)** = It is also known as specificity, is the amount of actual negative instances that the model correctly predicted as negative.
4. **FPR (False Positive Rate)** = It is defined as the amount of actual positive instances that the model incorrectly predicted as positive.

Precision, recall, accuracy and F1 score are also one of the most important metrics that are used for the evaluation of performance of a binary classification model, they are defined as follows:

Precision: It is a useful metric which shows that out of those predicted as positive, how many where actually positive [46].
Recall: It is also known as sensitivity or true positive rate and is defined as the amount of actual positive instances that the model accurately defined as positive.
Accuracy: It is the overall metric as to how often the classification is correct.
F1 score: It is basically defined as the harmonic mean of precision and recall, very small precision or recall will usually give a lower overall score, therefore, F1 score helps both the metrics to balance [47].

CHAPTER 4: RESULTS

Following are the results, objective wise.

1. **Dataset acquisition and pre-processing:** Data visualization is implemented graphically

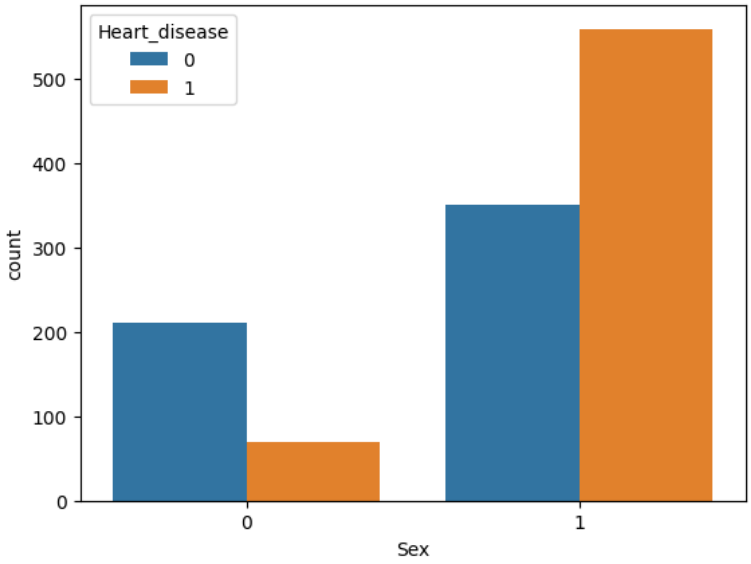


Figure 4 Graph between Sex and count

According to the above graph, men are more likely than women to suffer heart disease.

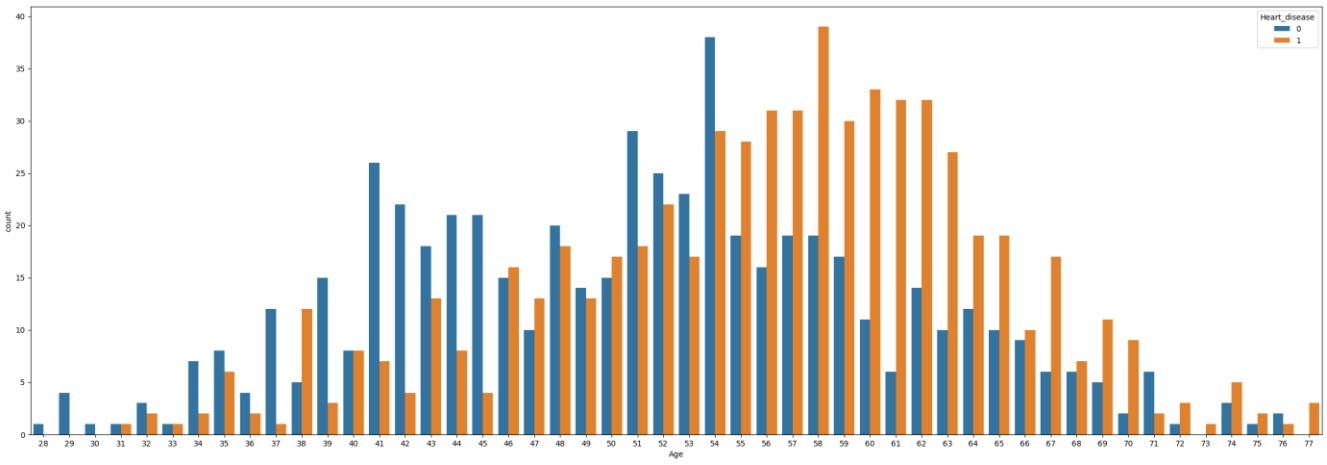


Figure 5 Graph between Age and count

One can conclude from the above graph that heart diseases are more common in adults over the age of 55 than in other age groups.

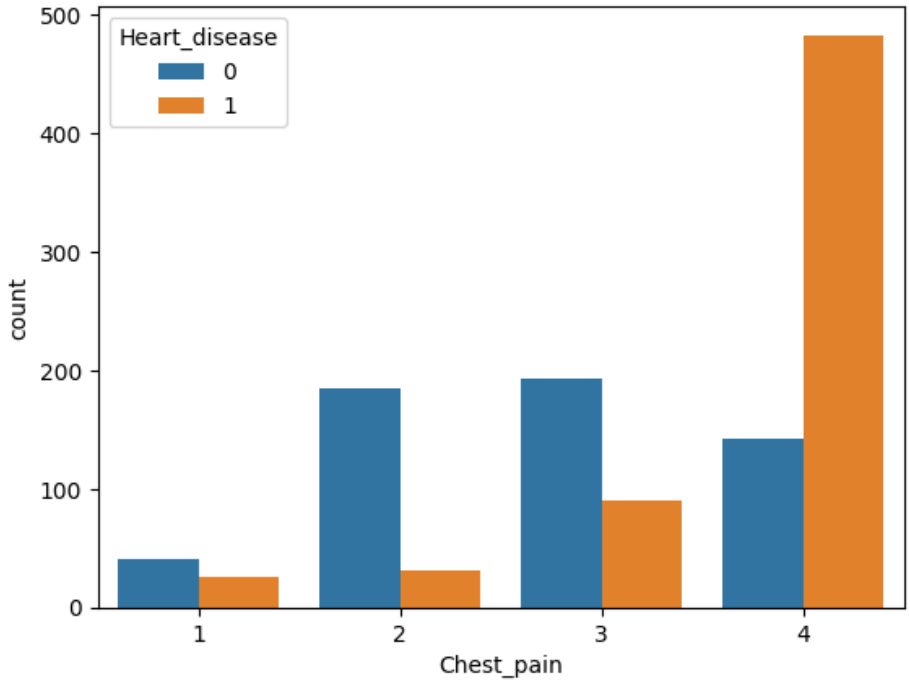


Figure 6 Graph between Chest pain and count

Above graph depicts that people with chest pain type 4 are very more likely to suffer from the heart disease than people from other types of chest pain.

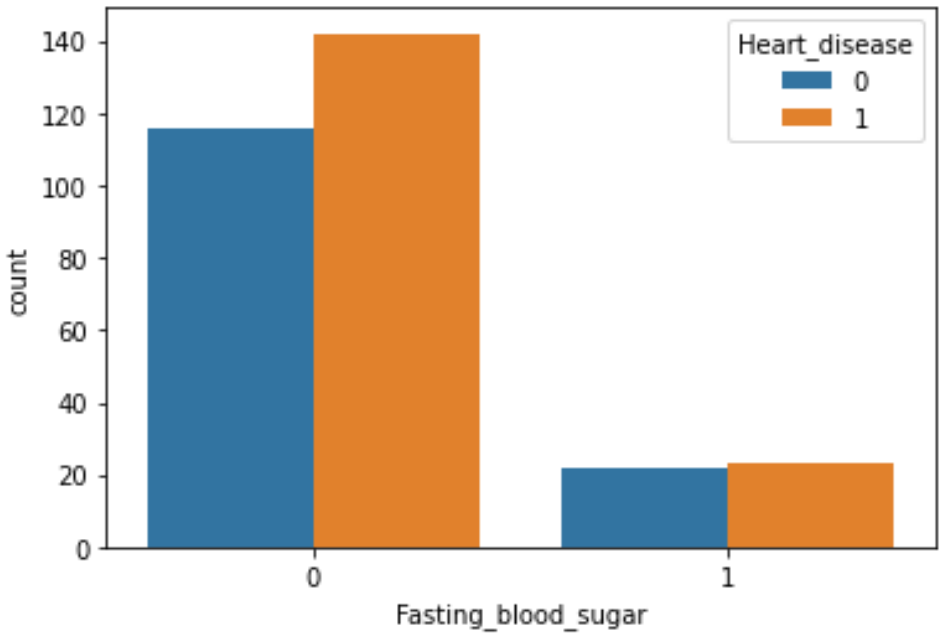


Figure 7 Graph between fasting blood sugar and count

People who have a fasting blood sugar level greater than 120 mg/dL are more susceptible to developing heart disease. However, it is not clear. This variable can be useful if we combine it with other variables.

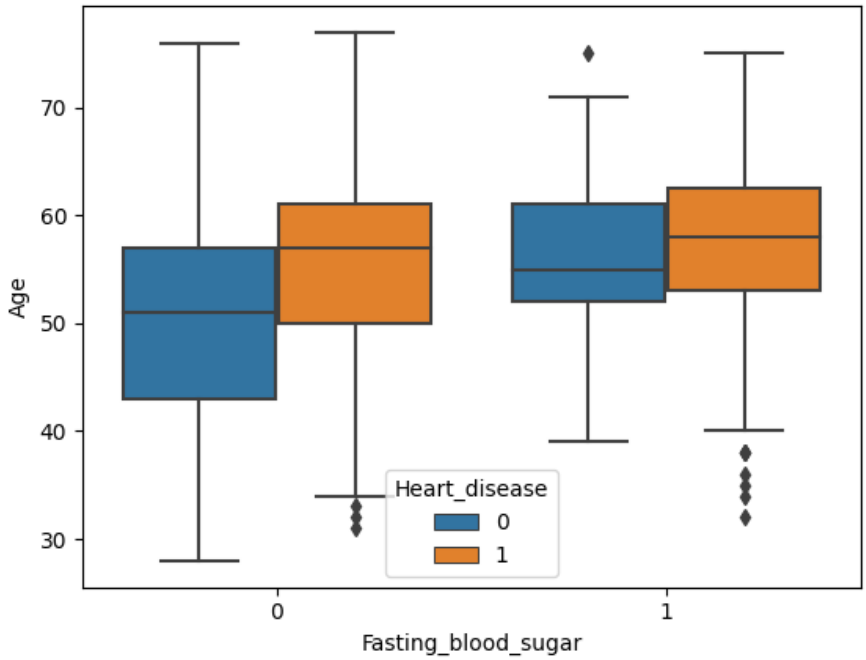


Figure 8 Graph between Fasting blood sugar and Age

We may now draw the conclusion that those who are older than 55 and have a fasting blood sugar level of more than 120 mg/dL are more likely to suffer heart disease.

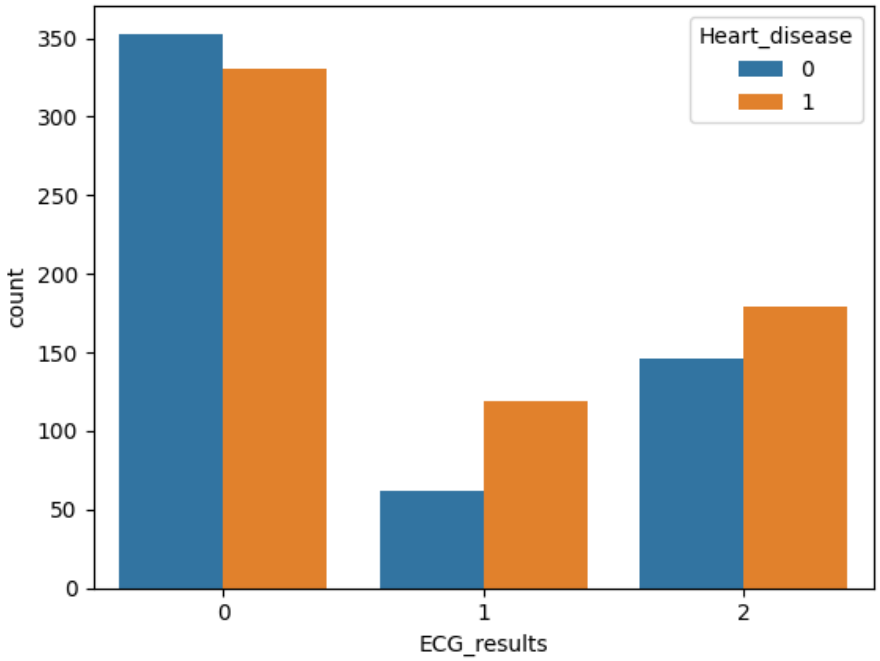


Figure 9 : Graph between ECG results and count

It shows that people having ventricular hypertrophy and abnormalities in T wave and ST segment have more chances of having heart disease (although we state it as it is not clear).

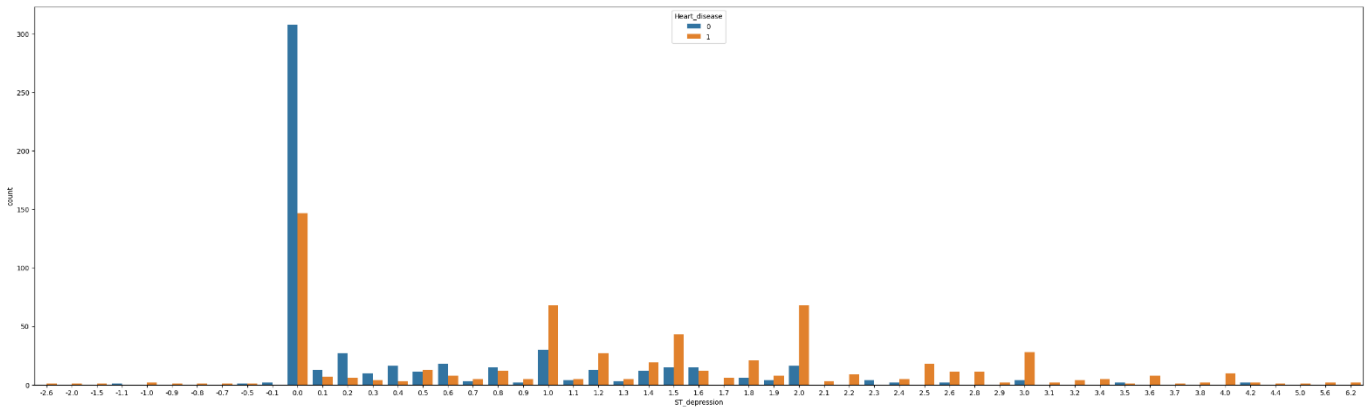


Figure 10 Graph between ST depression and count

As the ST Depression increases, probability of having heart disease increases.

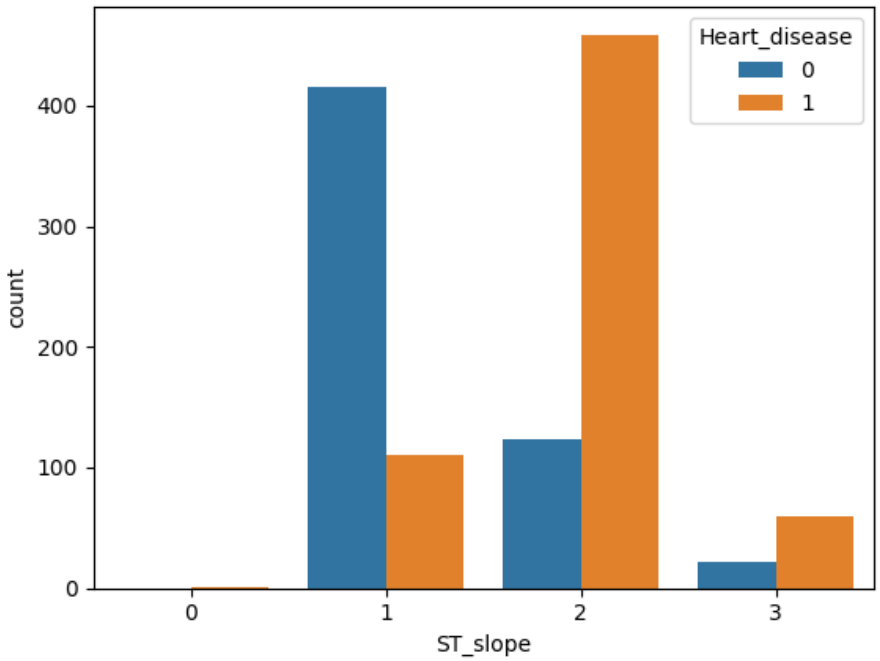


Figure 11 Graph between ST slope and count

When the ST slope is ascending, people have high probability of having heart diseases and when the slope is flat people are not having probability of heart diseases. This graph does not give us enough information about the descending part of the slope.

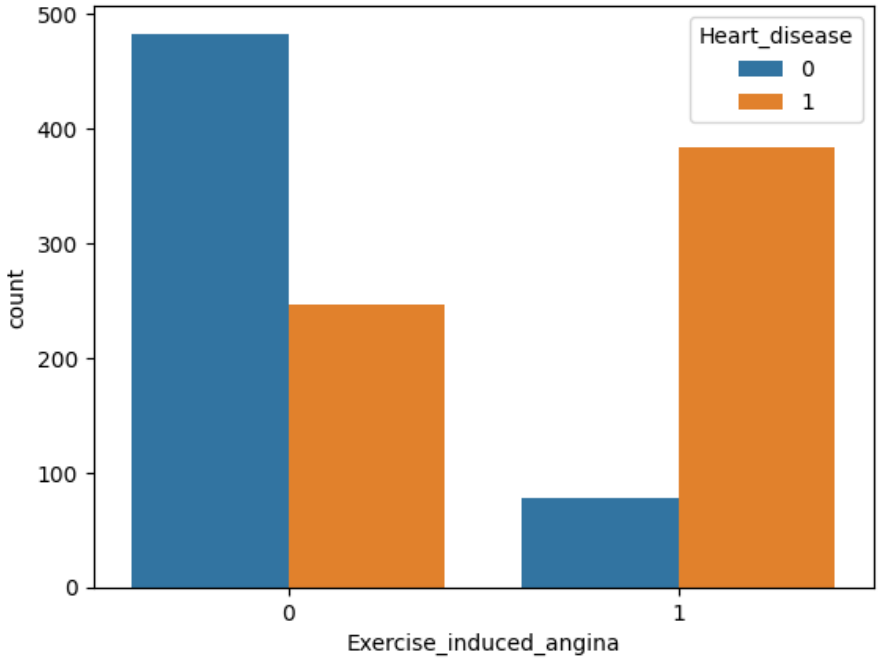


Figure 12 Graph between exercise induced angina and count

We can clearly see that people who experienced angina during exercise have good probability of having heart disease in comparison to those who did not.

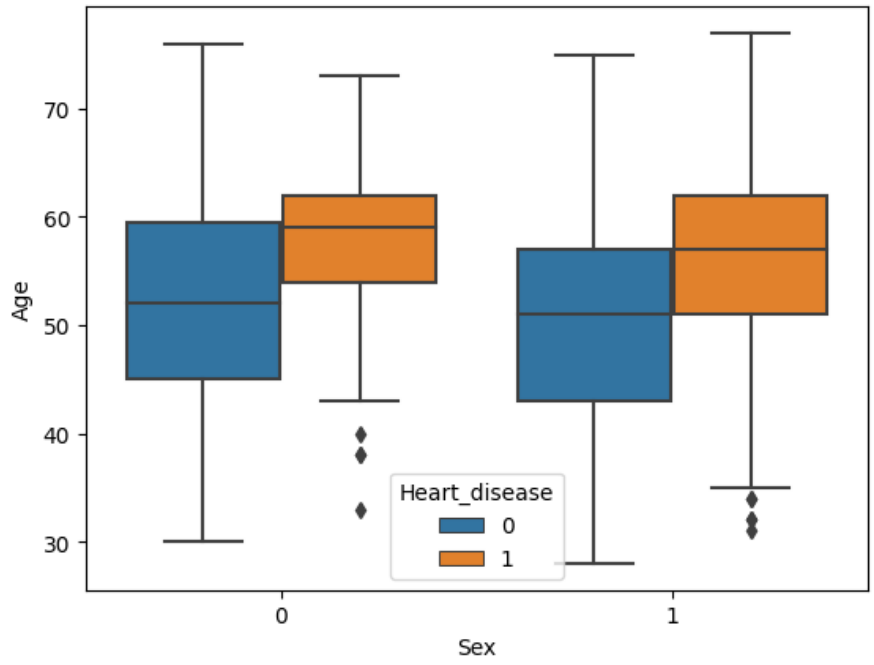


Figure 13 Graph between Sex and Age

We can conclude that males of age group 54-62 have highest chances of heart disease while female of age group 58-64 have highest chances of heart disease. As countplot has shown earlier, it also shows males have greater chances of having heart disease.

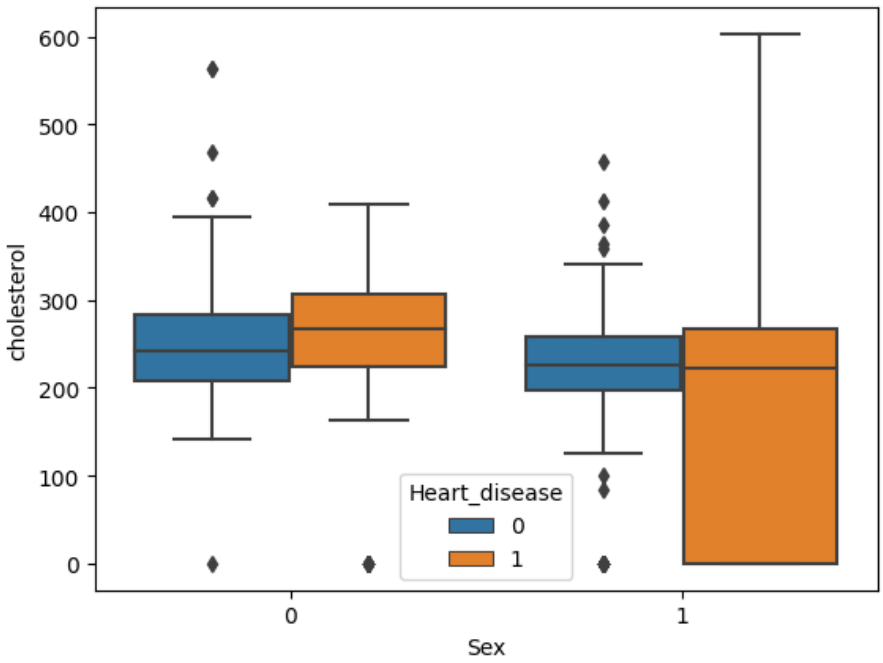


Figure 14 Graph between Sex and Cholesterol

People with high cholesterol have heart disease and females have higher cholesterol levels than males although females are less likely to suffer from heart disease than males.

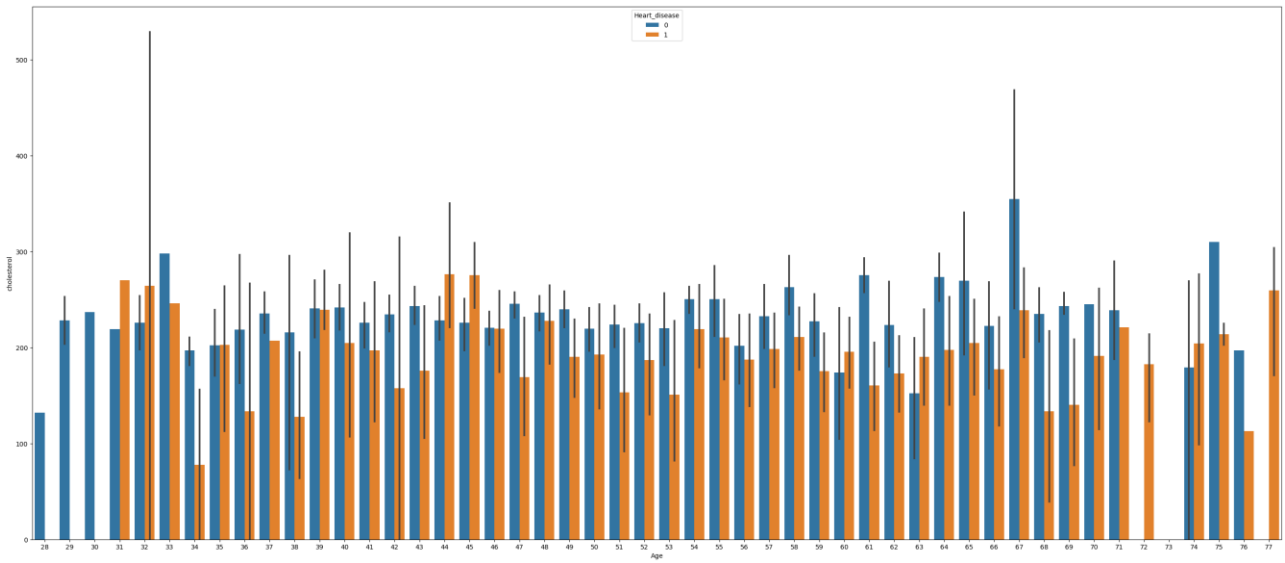


Figure 15 Graph between Age and Cholesterol

Cholesterol level is almost evenly distributed among the people of different age groups. It does not provide us any important information about the relation between cholesterol, age and heart disease. We can say that younger people with high cholesterol may not have heart disease as they do good amount of physical work.

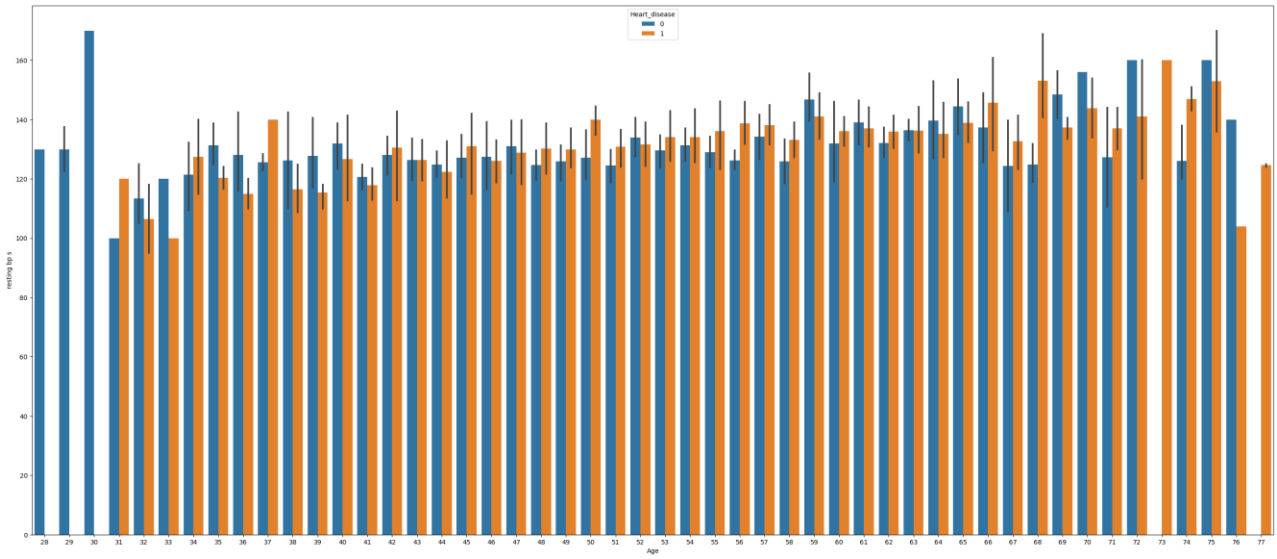


Figure 16 Graph between Age and Resting blood pressure

Very high blood pressure is shown by people having heart disease. Normal blood pressure is found in different age groups.

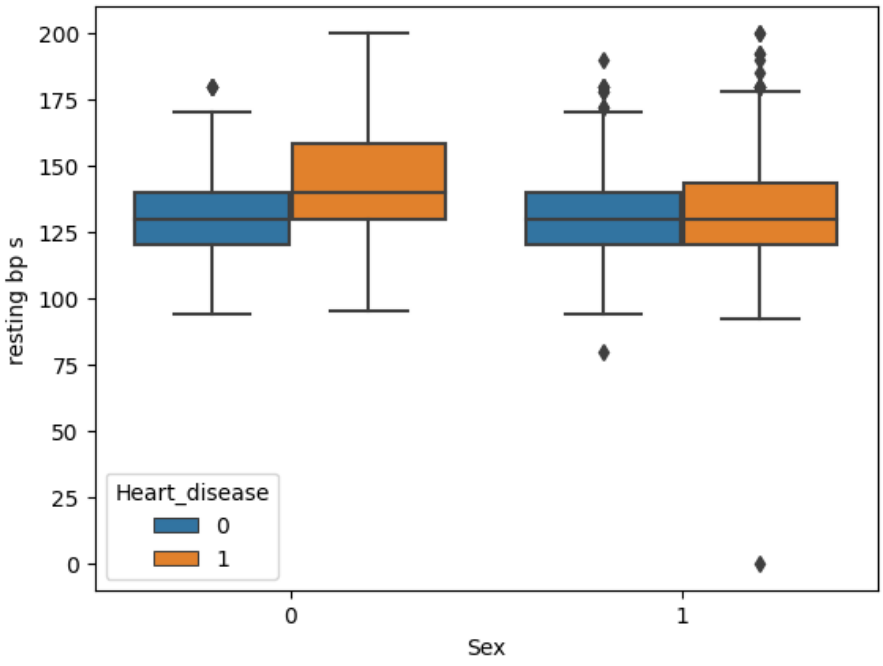


Figure 17 Graph between Sex and Resting blood pressure

Females are likely to have heart disease due to high blood pressure in comparison to males.

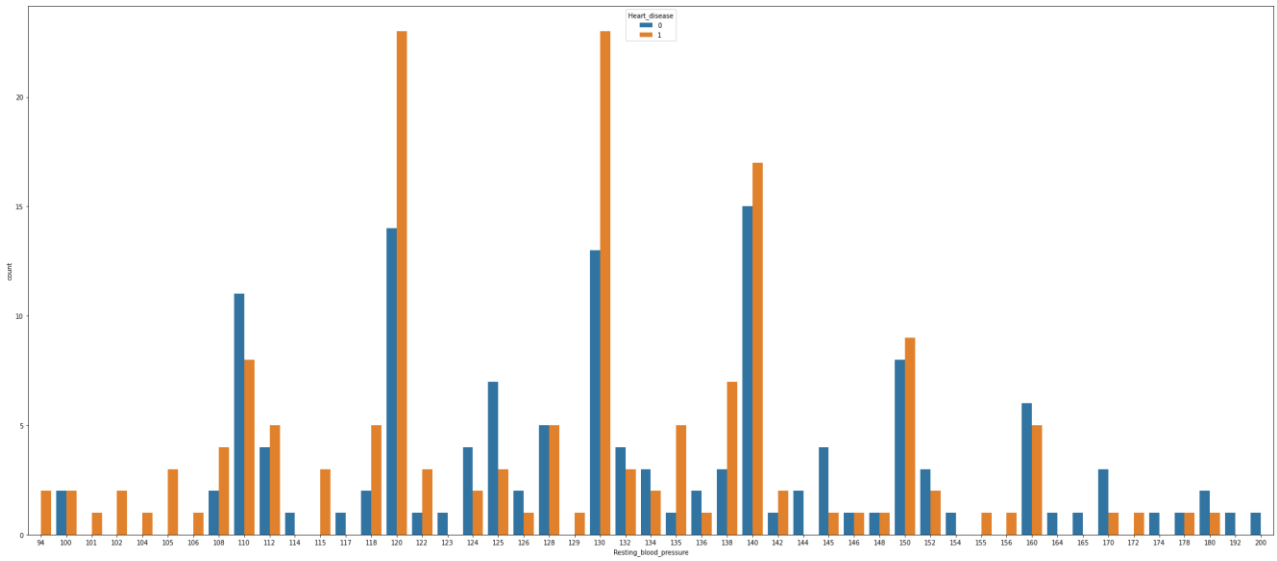


Figure 18 Graph between resting blood pressure and count

From the above figure, it can be concluded that high blood pressure is the indication of heart disease.

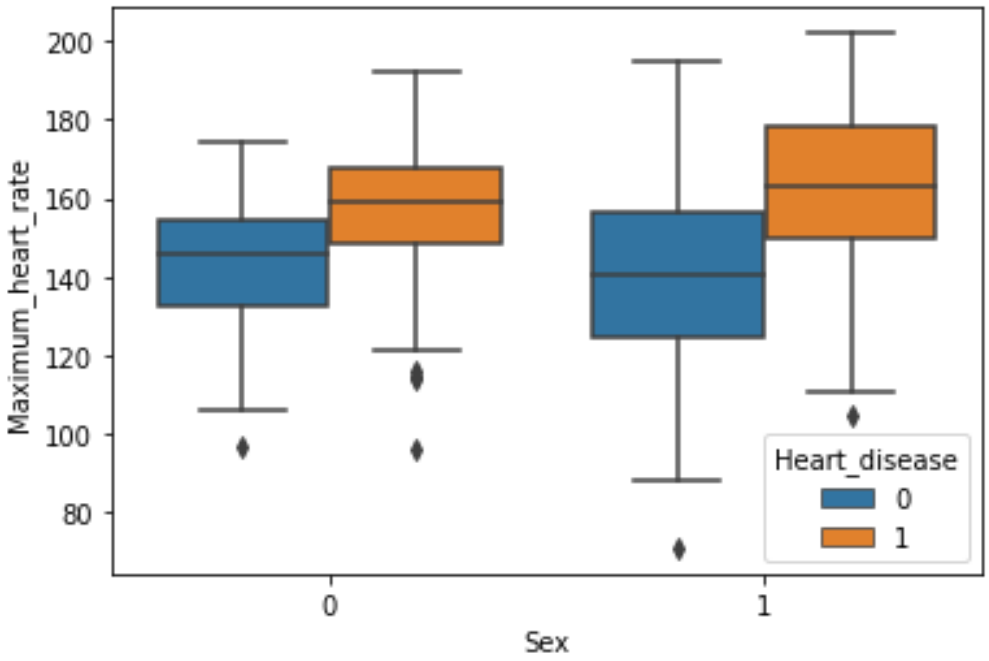


Figure 19 Graph between sex and Maximum heart rate

From the above graph, it can be concluded that greater the heart rate lesser the probability of having heart disease.

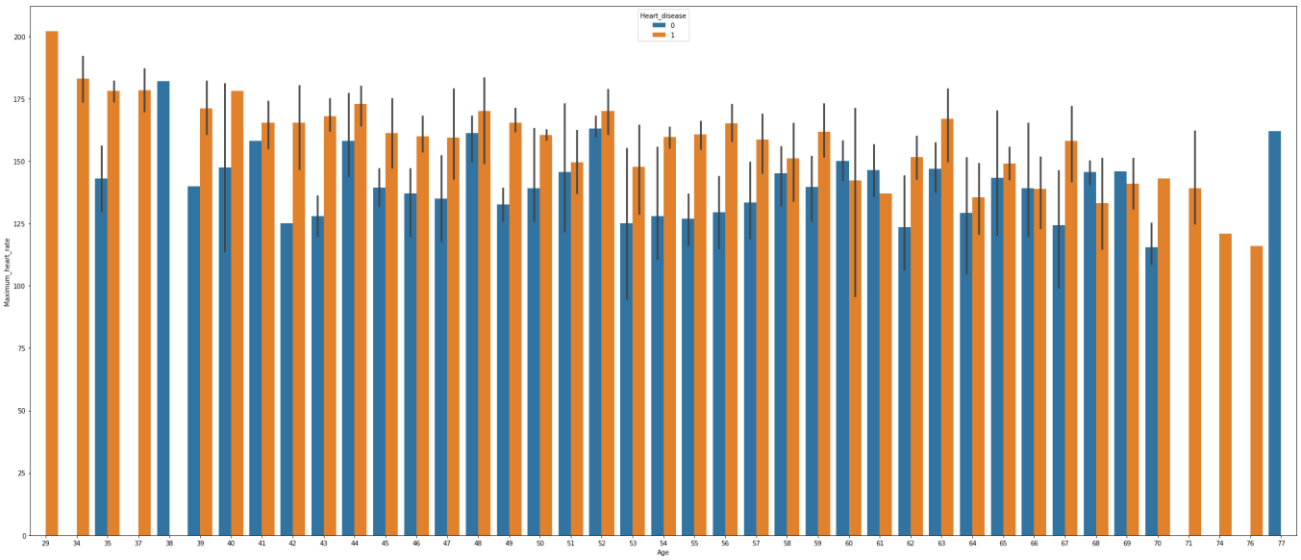


Figure 20 Graph between Age and Maximum heart rate

From the above graph in can be concluded that younger people have greater heart rate than older people.

2. Study and implementation of machine learning based classifiers to accurately detect CVDs

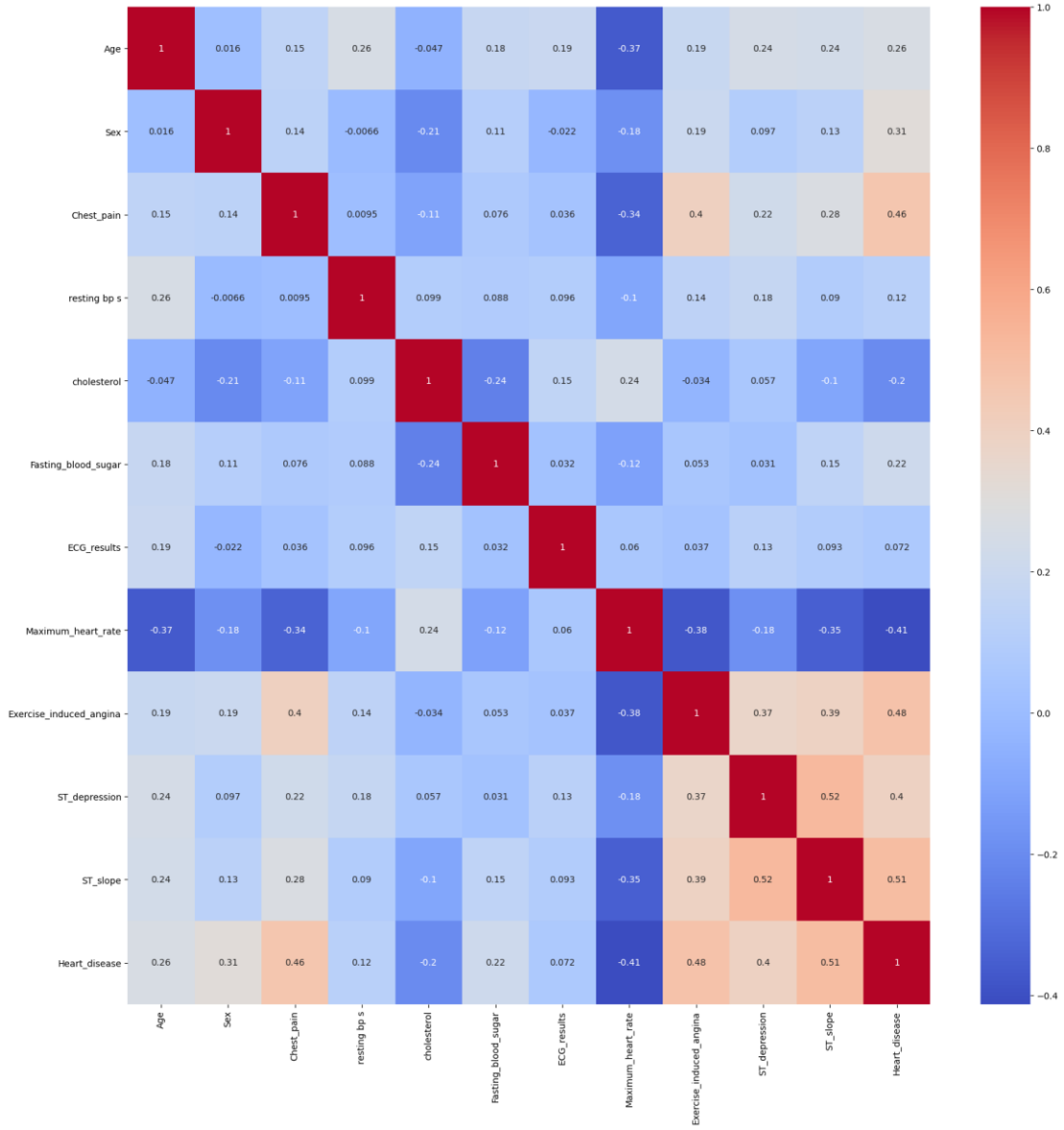


Figure 21 Heatmap showing correlations among all the features of the dataset.

Heatmap showing correlations between characteristics and correlated values. The correlation between two features and their correlated values is shown by all the coloured cells, with the colour of the cell representing the strength of the connection. A correlation value less than zero indicates a negative correlation, and a value of zero shows no correlation [15]. Heatmap provides us the insight that heart disease highly depends on chest pain and heart rate. ST slope and ECG results can also help us to predict heart disease.

Classification reports and confusion matrix:

1. Logistic Regression:

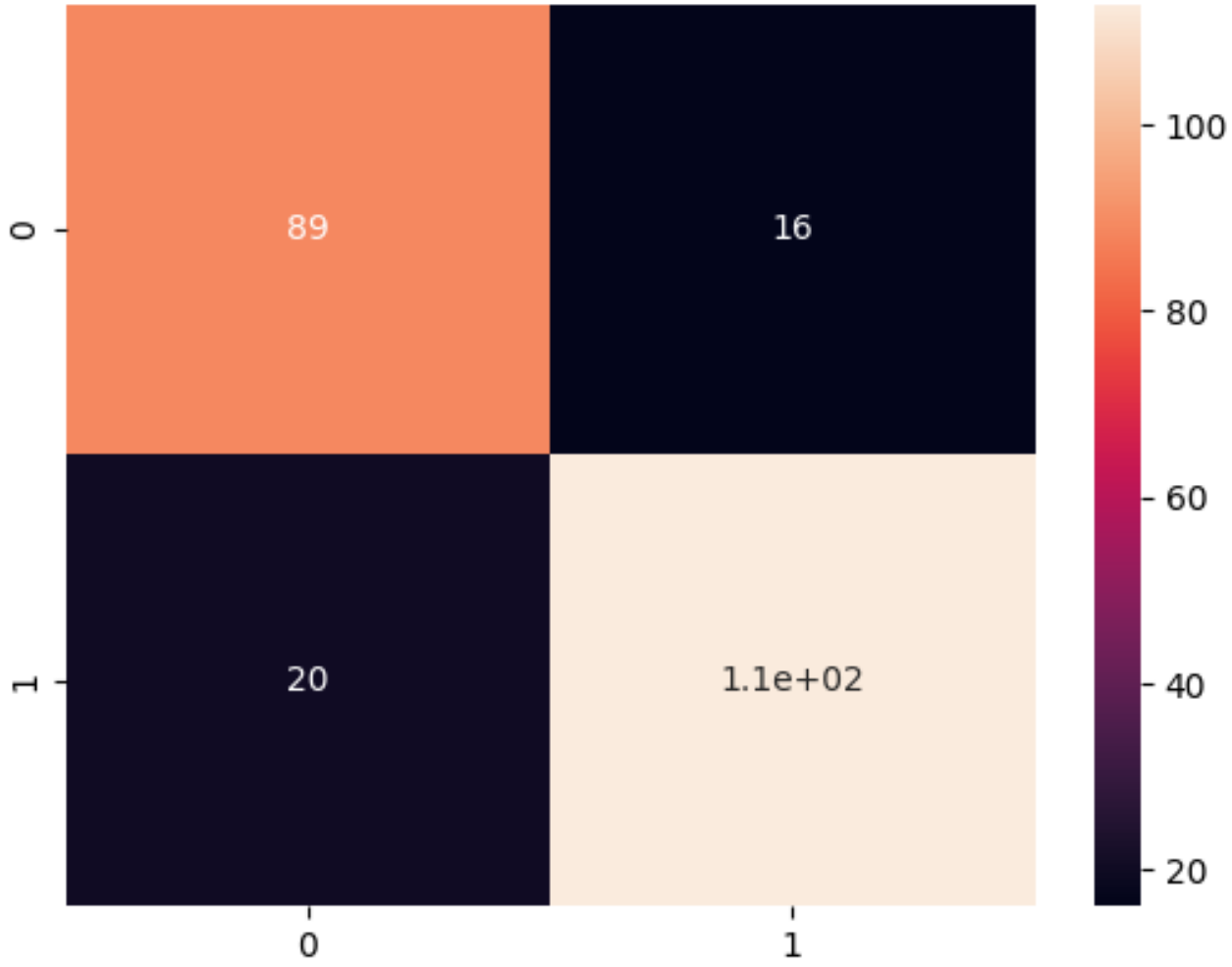


Figure 22: Logistic Regression confusion matrix

Table 1 Logistic Regression classification report

	Precision	Recall	F1 - score	support
0	0.82	0.85	0.83	105
1	0.88	0.85	0.86	133
Accuracy	-	-	0.85	238
Macro avg	0.85	0.85	0.85	238
Weighted avg	0.85	0.85	0.85	238

2. KNN

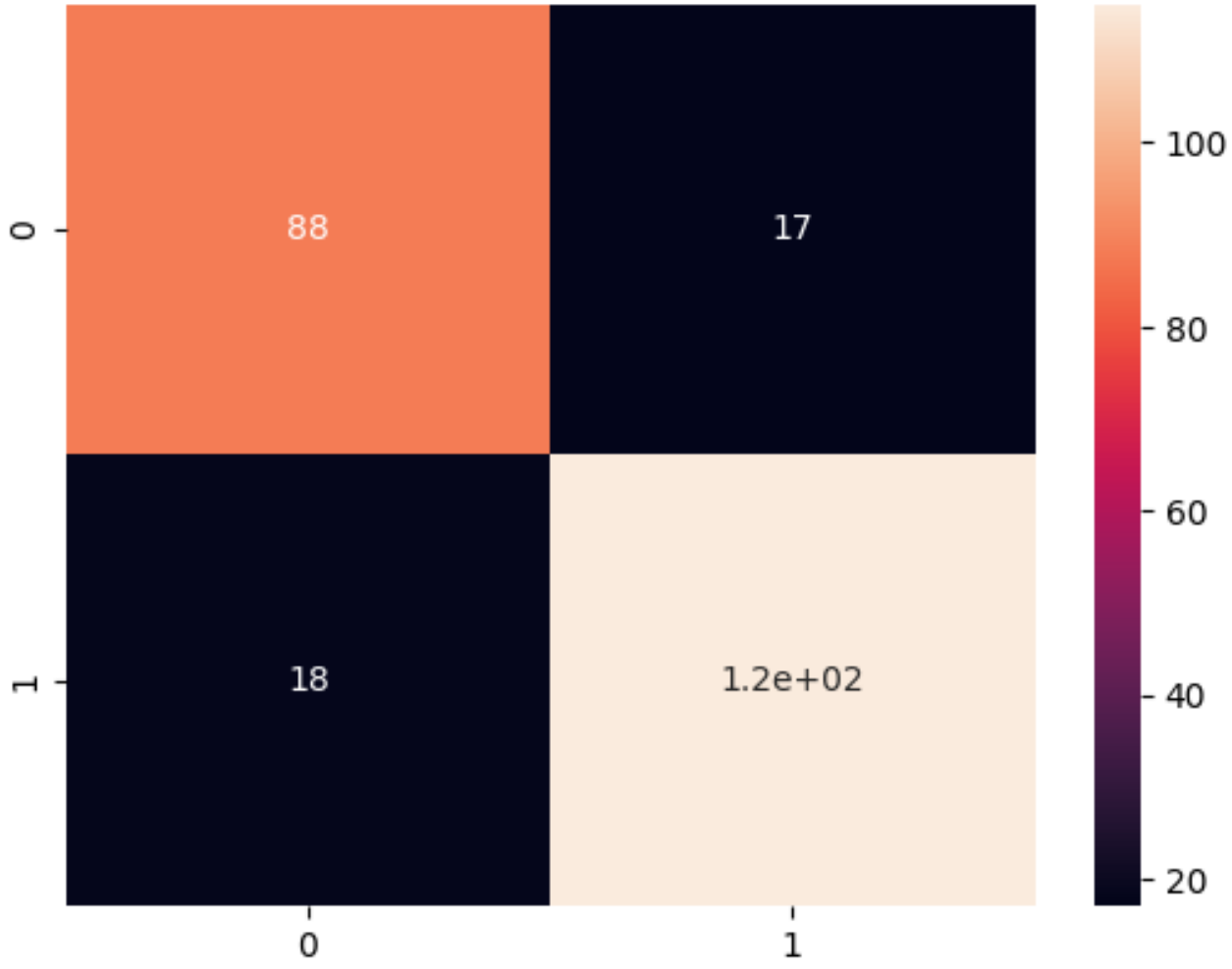


Figure 23: KNN confusion matrix

Table 2 KNN classification report

	Precision	Recall	F1 - score	support
0	0.83	0.84	0.83	105
1	0.87	0.86	0.87	133
Accuracy	-	-	0.85	238
Macro avg	0.85	0.85	0.85	238
Weighted avg	0.85	0.85	0.85	238

3. Random Forest Classifier

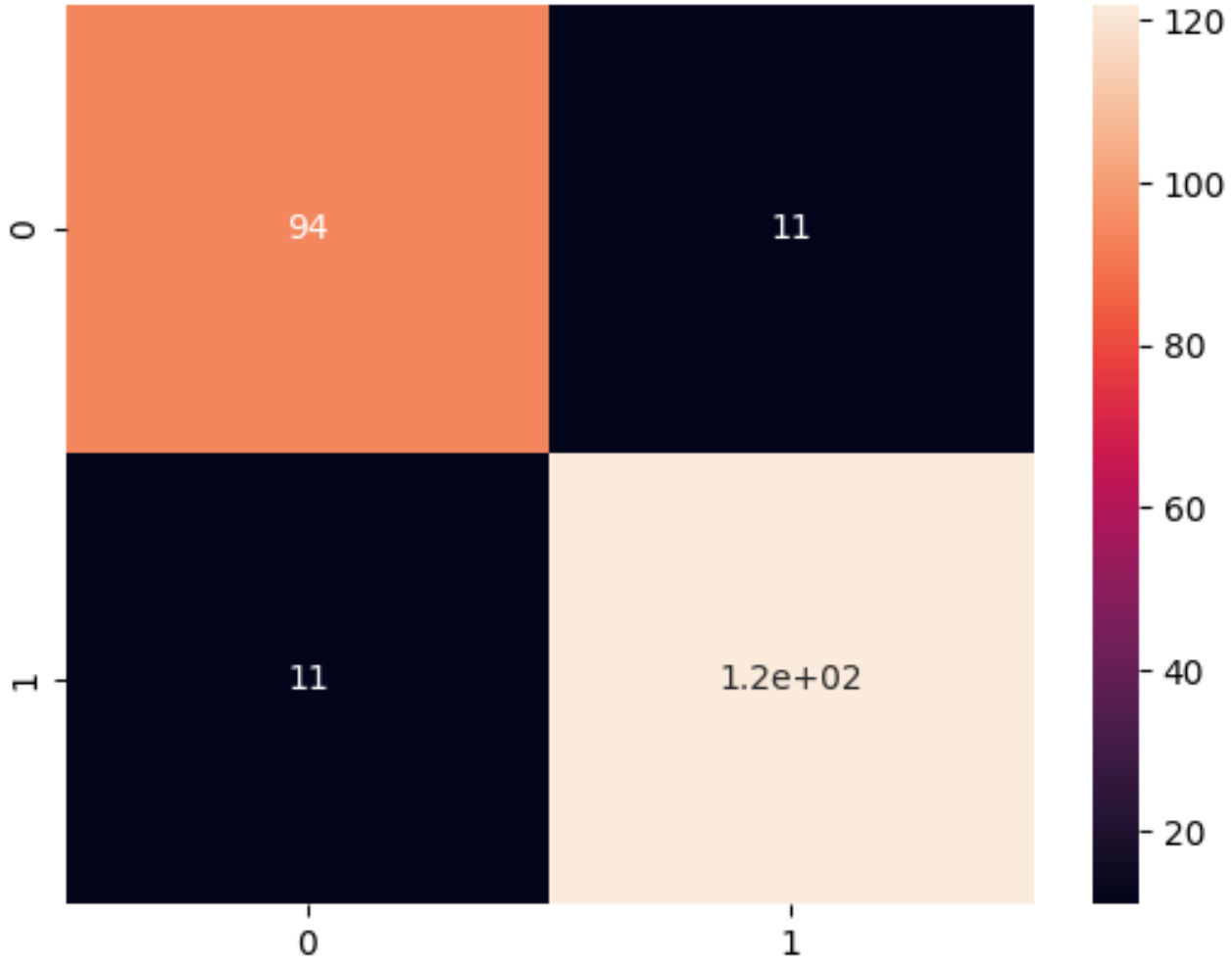


Figure 24: Random forest classifier confusion matrix

Table 3 Random Forest classifier classification report

	Precision	Recall	F1 - score	support
0	0.90	0.90	0.90	105
1	0.92	0.92	0.92	133
Accuracy	-	-	0.91	238
Macro avg	0.91	0.91	0.91	238
Weighted avg	0.91	0.91	0.91	238

4. SVM

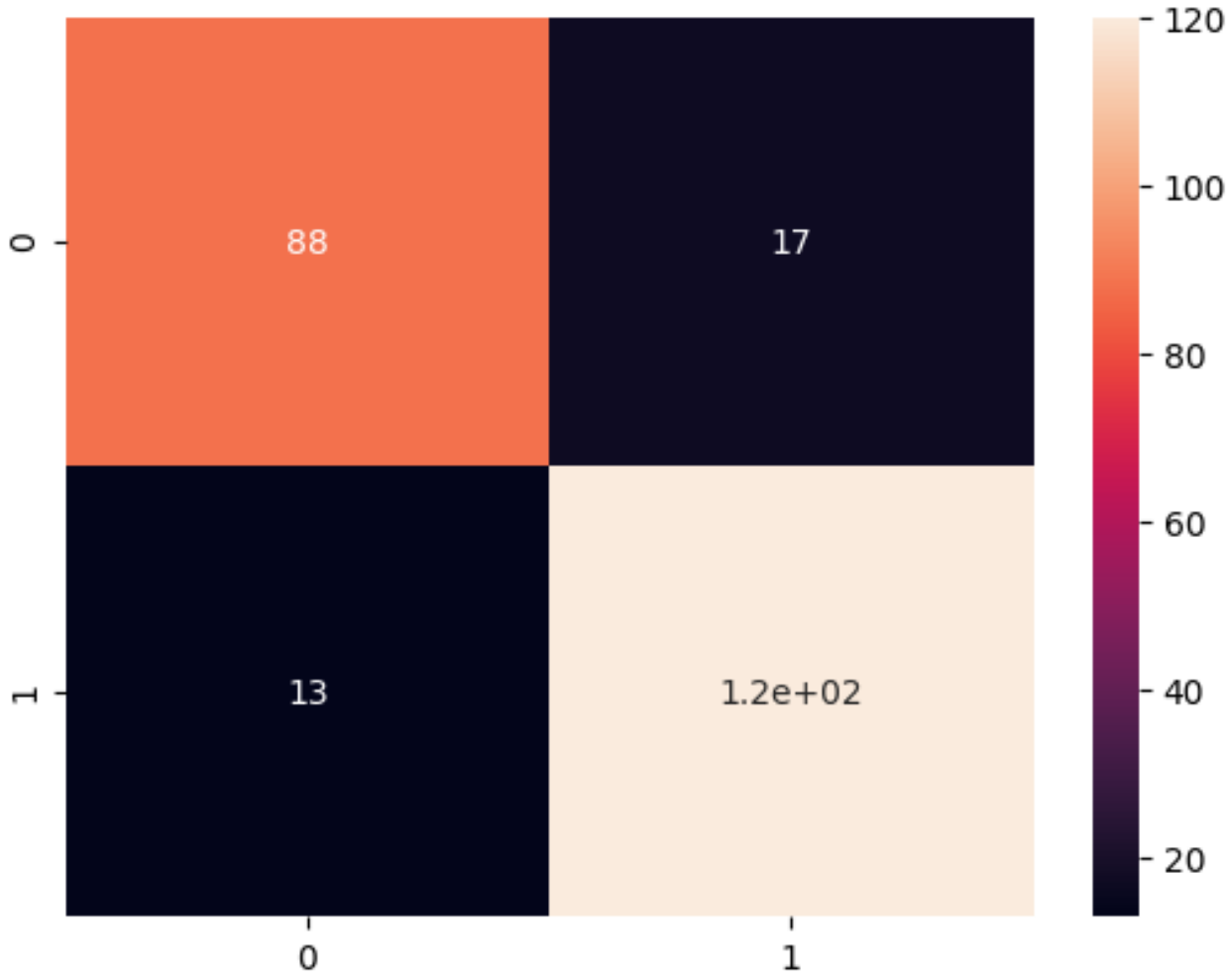


Figure 25: SVM confusion matrix

Table 4 SVM classification report

	Precision	Recall	F1 - score	support
0	0.87	0.84	0.85	105
1	0.88	0.90	0.89	133
Accuracy	-	-	0.87	238
Macro avg	0.87	0.87	0.87	238
Weighted avg	0.87	0.87	0.87	238

5. Decision Tree Classifier

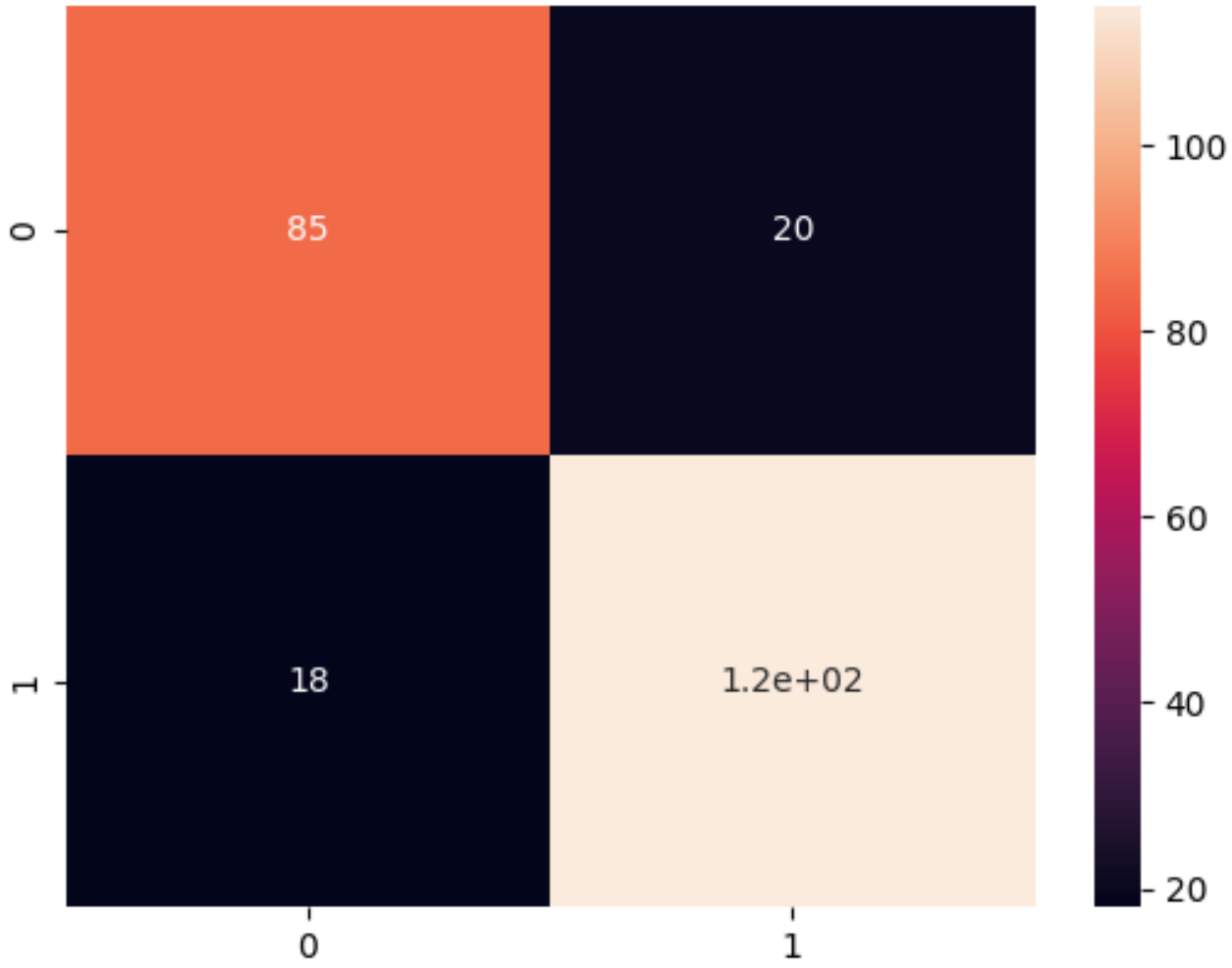


Figure 26: Decision Tree confusion matrix

Table 5 Decision Tree classifier classification report

	Precision	Recall	F1 - score	support
0	0.83	0.81	0.82	105
1	0.85	0.86	0.86	133
Accuracy	-	-	0.84	238
Macro avg	0.84	0.84	0.84	238
Weighted avg	0.84	0.84	0.84	238

6. Gaussian Naive Bayes

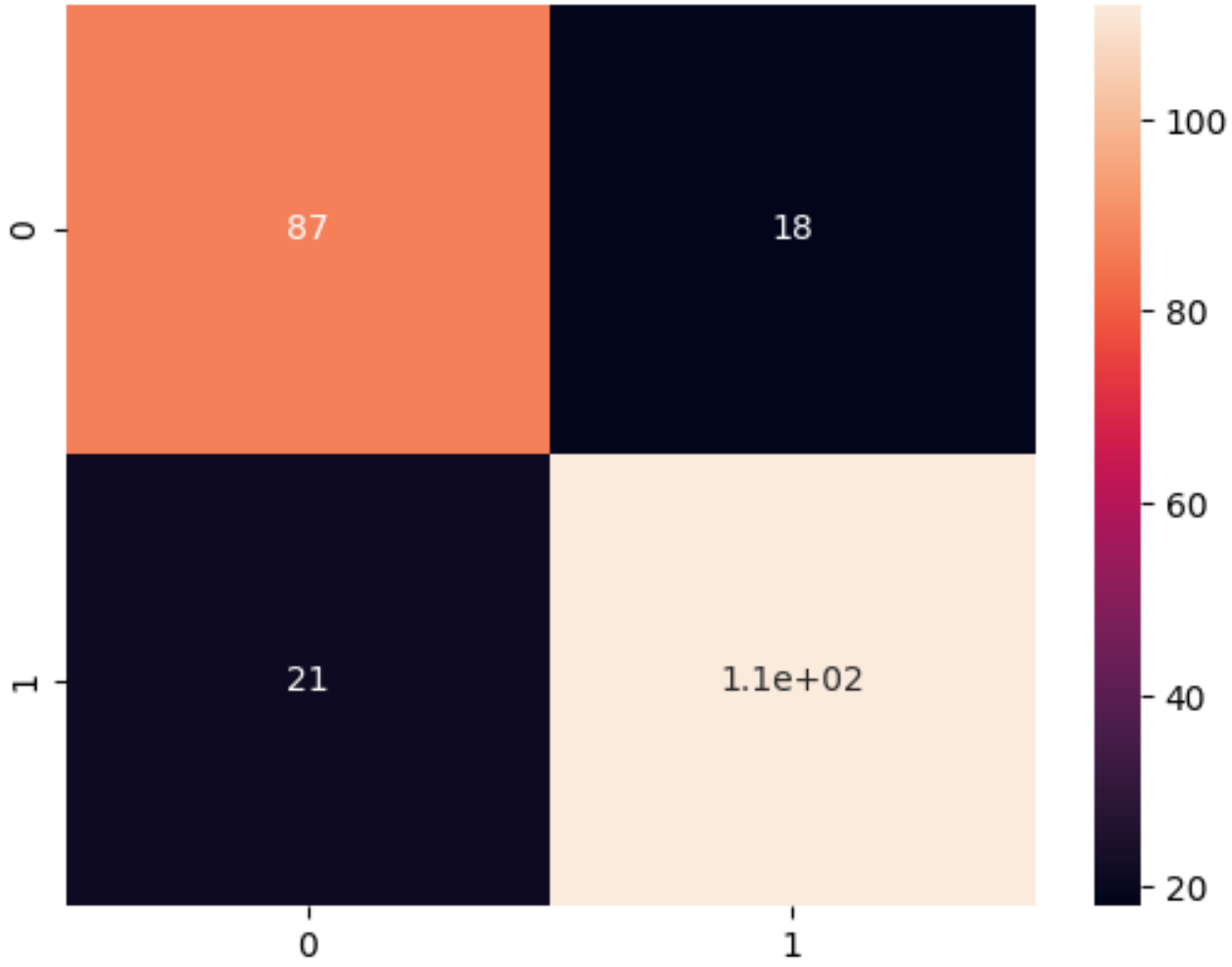


Figure 27: Gaussian Naive Bayes confusion matrix

Table 6 Gaussian Naive Bayes classification report

	Precision	Recall	F1 - score	support
0	0.81	0.83	0.82	105
1	0.86	0.84	0.85	133
Accuracy	-	-	0.84	238
Macro avg	0.83	0.84	0.83	238
Weighted avg	0.84	0.84	0.84	238

3. Comprehensive accuracy assessment summary and comparison between the classifiers to provide an overview of prediction models for the risk of CVD

Table 7: Comprehensive comparison between performance metrics of all models implemented for CVD

Algorithm	True Positive	True Negative	False Negative	False Positive	Accuracy	Precision	Recall	F1 -score
Logistic Regression	113	89	16	20	84.8%	0.85	0.87	0.86
KNN	115	88	17	18	85.2%	0.86	0.87	0.86
Random Forest Classifier	122	94	11	11	90.7%	0.92	0.92	0.92
SVM	120	88	17	13	87.3%	0.90	0.87	0.88
Decision Tree Classifier	115	85	20	18	84.0%	0.86	0.85	0.85
Gaussian Naive Bayes	112	87	18	21	83.6%	0.84	0.86	0.85

It is evident from the table above that the greatest value of accuracy, precision and recall is achieved through **Random Forest Classifier**. Therefore, it is the most efficient algorithm suited for prediction of cardiovascular diseases.

CHAPTER 5: CONCLUSION

Accurately detecting CVD is still a critical problem statement and requires real time data with expandable rich features. Application of ML based algorithms which can effectively extract features and predict the presence of CVD can play a crucial role in saving human lives and detection of abnormalities in heart conditions beforehand. Real time data from healthcare organizations and agencies needs to be procured and the available techniques can be implemented and compared to get the best accuracy. The likelihood that the model will correctly identify whether a specific person has heart disease or not increases with the use of more training data. In order to increase the prediction of accuracy in the early stages and enable the adoption of preventive measures as soon as possible, it is thought that only a marginal success is achieved in the development of a predictive model for patients with cardiovascular disease. As a result, combinational and more complex models are required in order to reduce the morality rate if the disease is discovered. As the accuracy of the model is significantly influenced by the dataset's quality, more hospitals should be encouraged to publish high-quality datasets (while protecting patient’s privacy) so that the researchers will have a trustworthy and genuine source to help them improve their models and obtain positive results that can help people benefit from and treat heart disease in its early stages. The results of this exhaustive study show that the Random Forest Classifier is the most efficient algorithm for predicting cardiovascular diseases with 90.756% accuracy, followed by SVM and KNN and Decision Tree Classifier. It can be concluded that there is a huge scope for ML algorithms in predicting cardiovascular diseases and there is still a need to research all the techniques available while fighting research gaps such as the capacity to handle high dimensional data, noisy datasets, maximum feature extraction and overfitting.

Limitations and future scope

- It should be noted that the amount of information on heart disease provided by this dataset was insufficient to fully address all issues, and that additional information and analysis are required to produce a reliable prediction method.
- The amount of processing power, memory, and storage needed to properly train and operate the model is referred to as the model's "computational power," which increases the cost and training time to use the model.
- This research can be conducted in the future using a variety of machine learning methodology combinations to improve prediction methods. Additionally, new feature-selection techniques can be created to obtain a broader understanding of the critical features and boost the accuracy of heart disease prediction.
- Deep learning is currently employed across all industries to achieve better outcomes. In the future, we'll work to implement additional deep learning algorithms to improve on Random Forest's performance.

References

[1] "Heart: Anatomy and Function," *Cleveland Clinic*. https://my.clevelandclinic.org/health/body/21704-heart (accessed May 02, 2023).

[2] Y. Jiang *et al.*, "Cardiovascular Disease Prediction by Machine Learning Algorithms Based on Cytokines in Kazakhs of China," *Clin. Epidemiol.*, vol. 13, p. 417, 2021, doi: 10.2147/CLEP.S313343.

[3] G. L. Team, "What is Machine Learning? Defination, Types, Applications, and more," *Great Learning Blog: Free Resources what Matters to shape your Career!*, Jan. 04, 2023. https://www.mygreatlearning.com/blog/what-is-machine-learning/ (accessed May 02, 2023).

[4] H. Singh, "Data Preprocessing," *Medium*, May 24, 2020. https://towardsdatascience.com/data-preprocessing-e2b0bed4c7fb (accessed May 02, 2023).

[5] J. Brownlee, "How to Choose a Feature Selection Method For Machine Learning," *MachineLearningMastery.com*, Nov. 26, 2019. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ (accessed May 02, 2023).

[6] A. Bhandari, "Feature Engineering: Scaling, Normalization, and Standardization (Updated 2023)," *Analytics Vidhya*, Apr. 03, 2020. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/ (accessed May 02, 2023).

[7] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020.

[8] "Logistic Regression for Machine Learning," *Capital One*. https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/ (accessed May 02, 2023).

[9] "What is a Decision Tree | IBM." https://www.ibm.com/topics/decision-trees (accessed May 02, 2023).

[10] "What are Neural Networks? | IBM." https://www.ibm.com/topics/neural-networks (accessed May 02, 2023).

[11] A. Javeed, M. A. Saleem, A. L. Dallora, L. Ali, J. S. Berglund, and P. Anderberg, "Decision Support System for Predicting Mortality in Cardiac Patients Based on Machine Learning," *Appl. Sci.*, vol. 13, no. 8, Art. no. 8, Jan. 2023, doi: 10.3390/app13085188.

[12] K. Mysiak, "Classification Metrics & Thresholds Explained," *Medium*, Aug. 07, 2020. https://towardsdatascience.com/classification-metrics-thresholds-explained-caff18ad2747 (accessed May 02, 2023).

[13] J. A. Damen *et al.*, "Prediction models for cardiovascular disease risk in the general population: systematic review," *bmj*, vol. 353, 2016.

[14] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, p. 012072, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012072.

[15] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput. Biol. Med.*, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

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

[17] V. V. Ramalingam, A. Dandapath, and M. Raja, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, p. 684, Mar. 2018, doi: 10.14419/ijet.v7i2.8.10557.

[18] V. V. Ramalingam, A. Dandapath, and M. Raja, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, p. 684, Mar. 2018, doi: 10.14419/ijet.v7i2.8.10557.

[19] G. C. Siontis, I. Tzoulaki, K. C. Siontis, and J. P. Ioannidis, "Comparisons of established risk prediction models for cardiovascular disease: systematic review," *Bmj*, vol. 344, 2012.

[20] J. Patel, S. Tejalupadhyay, and S. Patel, *Heart Disease prediction using Machine learning and Data Mining Technique*. 2016. doi: 10.090592/IJCSC.2016.018.

[21] S. Nikhar and A. M. Karandikar, "Prediction of Heart Disease Using Machine Learning Algorithms," vol. 2, no. 6.

[22] Hassan 1st University, Y. Khourdifi, M. Bahaj, and Hassan 1st University, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 242–252, Feb. 2019, doi: 10.22266/ijies2019.0228.24.

[23] N. B. Amma, "Cardiovascular disease prediction system using genetic algorithm and neural network," in *2012 International Conference on Computing, Communication and Applications*, IEEE, 2012, pp. 1–5.

[24] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm," *Procedia Technol.*, vol. 10, pp. 85–94, 2013, doi: 10.1016/j.protcy.2013.12.340.

[25] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[26] J. Kiran, N. Debbarma, and S. Ganjala, "Heart Disease Prediction Using Machine Learning," in *Evolution in Computational Intelligence*, Springer, Singapore, 2023, pp. 263–272. doi: 10.1007/978-981-19-7513-4_24.

[27] S. M. Kharabsheh, A. Al-Sugair, J. Al-Buraiki, and J. Farhan, "Overview of Exercise Stress Testing," *Ann. Saudi Med.*, vol. 26, no. 1, p. 1, Feb. 2006, doi: 10.5144/0256-4947.2006.1.

[28] "What is NumPy? — NumPy v1.23 Manual." https://numpy.org/doc/stable/user/whatisnumpy.html (accessed Nov. 29, 2022).

[29] "pandas - Python Data Analysis Library." https://pandas.pydata.org/ (accessed Nov. 29, 2022).

[30] "Matplotlib — Visualization with Python." https://matplotlib.org/ (accessed Nov. 29, 2022).

[31] M. Waskom, "seaborn: statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[32] "What is Logistic regression? | IBM." https://www.ibm.com/in-en/topics/logistic-regression (accessed Nov. 29, 2022).

[33] S. Swaminathan, "Logistic Regression — Detailed Overview," *Medium*, Jan. 18, 2019. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc (accessed Nov. 29, 2022).

[34] "*k*-nearest neighbors algorithm," *Wikipedia*. Nov. 10, 2022. Accessed: Nov. 29, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=1121170714

[35] "K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint," *www.javatpoint.com*. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (accessed Nov. 29, 2022).

[36] "sklearn.ensemble.RandomForestClassifier," *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed Nov. 30, 2022).

[37] "Introduction to Random Forest in Machine Learning," *Engineering Education (EngEd) Program | Section*. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/ (accessed Nov. 30, 2022).

[38] "Support Vector Machine (SVM) Algorithm - Javatpoint." https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm (accessed Nov. 30, 2022).

[39] "Support vector machine," *Wikipedia*. Oct. 19, 2022. Accessed: Nov. 30, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=1116926893

[40] "Machine Learning Decision Tree Classification Algorithm - Javatpoint," *www.javatpoint.com*. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (accessed Nov. 30, 2022).

[41] "Decision Tree," *GeeksforGeeks*, Oct. 16, 2017. https://www.geeksforgeeks.org/decision-tree/ (accessed Nov. 30, 2022).

[42] S. Hrouda-Rasmussen, "(Gaussian) Naive Bayes," *Medium*, May 07, 2021. https://towardsdatascience.com/gaussian-naive-bayes-4d2895d139a (accessed Nov. 30, 2022).

[43] S. K. J. and G. S., “Prediction of Heart Disease Using Machine Learning Algorithms,” in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, Apr. 2019, pp. 1–5. doi: 10.1109/ICIICT1.2019.8741465.

[44] V. Jayaswal, “Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score,” *Medium*, Sep. 15, 2020. https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262 (accessed Nov. 30, 2022).

[45] “Confusion matrix,” *Wikipedia*. Apr. 07, 2023. Accessed: May 03, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=1148699071

[46] Naveen, “What is precision, Recall, Accuracy and F1-score?,” *Nomidl*, Feb. 27, 2022. https://www.nomidl.com/machine-learning/what-is-precision-recall-accuracy-and-f1-score/ (accessed May 03, 2023).

[47] T. Kanstrén, “A Look at Precision, Recall, and F1-Score,” *Medium*, May 19, 2021. https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec (accessed May 03, 2023).