Mini Project Report on

HEART DISEASE PREDICTION BY MACHINE LEARNING

Submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Student Name:	University Roll no.
Ashish Upadhyay	2017474

Under the Mentorship of

Dr. Surendra Kumar Shukla
Designation

Department of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, Uttarakhand
July-2023

Graphic Era
Deemed to be
University
DEHRADUN

CANDIDATE’S DECLARATION

I hereby certify that the work which is being presented in the project report entitled **“Heart Disease Prediction by Machine Learning”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Surendra Kumar Shukla, Designation**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Ashish Upadhyay University Roll no: 2017474

Table of Contents

Chapter 1

Introduction

1.1 Introduction

Heart disease is a widespread and significant health concern that affects millions of people worldwide. It is a leading cause of mortality and poses a considerable burden on healthcare systems globally. The ability to accurately predict the occurrence of heart disease can greatly aid in its prevention and timely intervention. In recent years, machine learning algorithms have emerged as powerful tools for analyzing large datasets and extracting meaningful patterns that can help in predicting heart disease. These advanced computational methods have the potential to empower the field of cardiovascular medicine by providing reliable predictions and assisting medical professionals in making informed decisions.

This topic explores the exciting intersection between machine learning and heart disease prediction. By leveraging the available medical data, researchers and data scientists are developing sophisticated models and algorithms to predict the likelihood of heart disease in individuals. Machine learning algorithms can analyze various factors, including demographic information, lifestyle choices, medical history, and diagnostic test results, to generate accurate predictions. By identifying high-risk individuals early on, preventative measures can be implemented, such as lifestyle modifications, medication, or targeted interventions, to reduce the chances of developing heart disease or to mitigate its impact.

In this context, machine learning algorithms offer several advantages over traditional statistical approaches. These algorithms can automatically learn from data without being explicitly programmed, allowing them to discover and observes complex relationships and patterns which otherwise could’ve been missed by traditional methods. Additionally, machine learning models can continuously improve their predictions as they process more data, making them highly adaptable and robust. Moreover, the integration of machine learning with electronic health records and wearable devices has the potential to enable real-time monitoring and personalized interventions, leading to more effective and efficient healthcare delivery.

Despite the promising potential of machine learning in heart disease prediction, there are challenges that need to be addressed. Data quality, privacy concerns, interpretability of models, and generalizability across diverse populations are among the key issues that researchers and practitioners face. Overcoming these challenges requires collaborative efforts between medical professionals, data scientists, and policymakers to ensure the ethical and responsible use of machine learning techniques in healthcare.

As we know that there is no perfect algorithm that fits all the data for every need. So, a variety of algorithms were used in this project to better analyze the accuracy of models based on each of the algorithms.

The algorithms explored are:

1) Logistic Regression
2) Random Forest
3) K Nearest Neighbor
4) Support Vector Machine

1.1.1 Logistic Regression

It is used for binary classification problems where the outcome variable is categorical with two classes.

1.1.2 Random Forest

It is suitable for both classification and regression problems. It is used when the dataset is complex and has high dimensional data interactions between the variables.

1.1.3 K Nearest Neighbor

It also works on both classification and regression problems. It is used when the decision boundary is complex and not easily characterized by a mathematical function.

1.1.4 Support Vector Machines

Used to deal with complex, non-linear relationship between variables. It aims to find a hyperplane that separates the classes while maximizing the margin between the two.

1.2 System Requirements

- Internet Connection
- A PC or Laptop
- Dataset (from Kaggle)
- Google Colaboratory

Chapter 2

Literature Survey

The following work was done in this field by using the above-mentioned algorithms:

2.1 Logistic Regression

A study by Dey et al. (2017) used logistic regression to predict the risk of coronary artery disease based on clinical and demographic factors. The model showed promising accuracy and identified significant predictors such as age, cholesterol levels, and smoking status.

A study conducted by Zhang et al. (2018), logistic regression was employed to predict the occurrence of heart failure in patients with hypertension. The model incorporated features such as blood pressure, body mass index, and laboratory test results, achieving good predictive performance.

2.2 Random Forest

In a research study by Harikrishnan et al. (2019), a random forest algorithm was utilized to predict the presence of heart disease based on clinical and echocardiographic parameters. The model demonstrated high accuracy and outperformed other machine learning algorithms in the study.

Rizvi et al. (2020) employed random forest to predict cardiovascular events in patients with diabetes. The model incorporated clinical, biochemical, and electrocardiographic data, achieving good predictive accuracy and providing valuable insights into risk stratification.

2.3 K-Nearest Neighbors (KNN)

A study by Liu et al. (2017) utilized KNN to predict the presence of coronary artery disease based on features extracted from coronary computed tomography angiography images. The KNN model achieved high accuracy in detecting the disease and showed potential for non-invasive diagnosis.

In another investigation by Anwar et al. (2020), KNN was used to predict the risk of cardiovascular diseases based on lifestyle factors and medical history. The study highlighted the effectiveness of KNN in identifying high-risk individuals and supporting preventive interventions.

2.4 Support Vector Machines (SVM)

A research study by Alizadehsani et al. (2013) employed SVM to predict the presence of heart disease based on features extracted from electrocardiogram signals. The SVM model achieved high accuracy and showed potential for early detection of cardiac abnormalities.

In a study by Wu et al. (2020), SVM was utilized to predict major adverse cardiac events in patients with coronary artery disease. The model incorporated clinical and laboratory data, demonstrating good predictive performance and assisting in risk stratification.

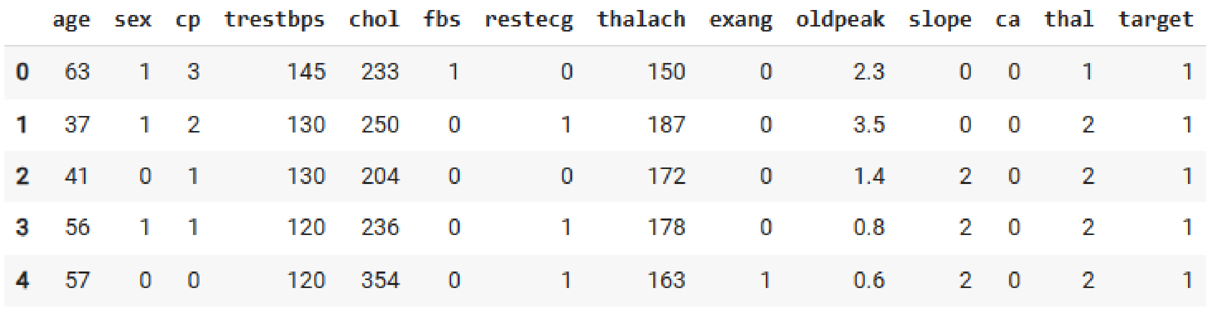
Chapter 3

Methodology

The following steps were followed in making this project:

3.1 Dataset Description

The dataset used is publicly available patient data. It contains data such as – age of the patient (years), sex of the patient (male/female), cholesterol levels (mg/dl), blood pressure (mmHg), type of chest pain (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic), maximum heart rate achieved, exercise induced angina, results (disease/healthy), and other clinical test data.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

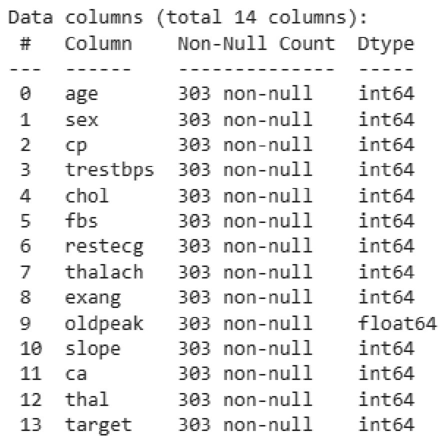
Fig 3.1 Overview of the dataset used

3.2 Data Preprocessing

It is a vital step in machine learning, as the dataset provided can contain missing or null values. We need to find and handle them so that the data that is being fed to the model is noiseless. The following steps are involved in data preprocessing:

3.2.1 Handling Missing or Null Values

Missing or null values are handled by replacing them with the mean, median or mode of the neighbor elements, deleting the entire tuple, etc. whichever is demanded by the given situation



```
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
```

Fig 3.2 Checking for NULL values in the dataset

3.3 Feature Selection

Choosing right features for training the model is essential as not all the features present in the dataset are truly required by the model. One way to do that is by find correlation among different variable in the dataset and then choosing the ones that are the most closely correlated to the result.

3.4 Developing the model

As no single algorithm can be used to fit all the data for every use, four algorithms were tested for their accuracy and precision.

3.4.1 Logistic Regression

Logistic Regression is a widely used algorithm for binary classification and works well if the dataset has smaller number of values and is linear. It is suitable where the number of computational requirements is low.

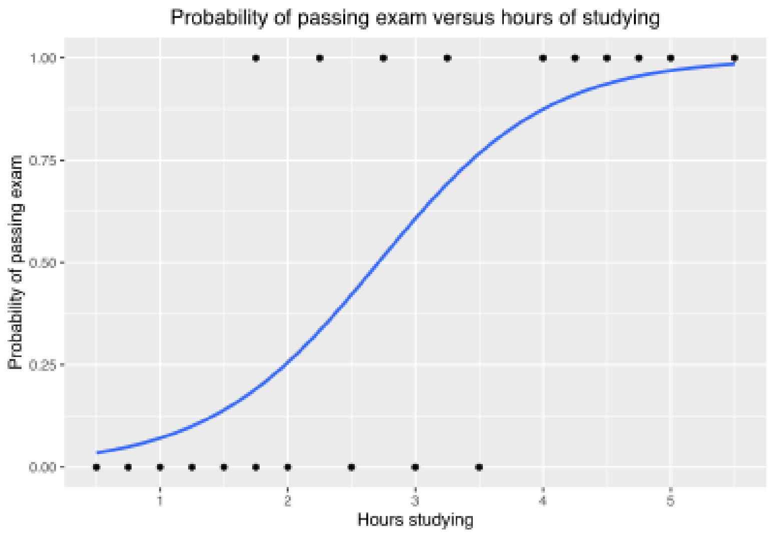


Fig 3.3 Graph of Logistic Regression

3.4.2 Random Forest

It is used for large and complex datasets where there is data of high dimensionality. It uses multiple decision trees and is less prone to overfitting, which makes this algorithm very robust. It can handle large number of input features and can automatically select important features thus reducing the work of feature engineering.

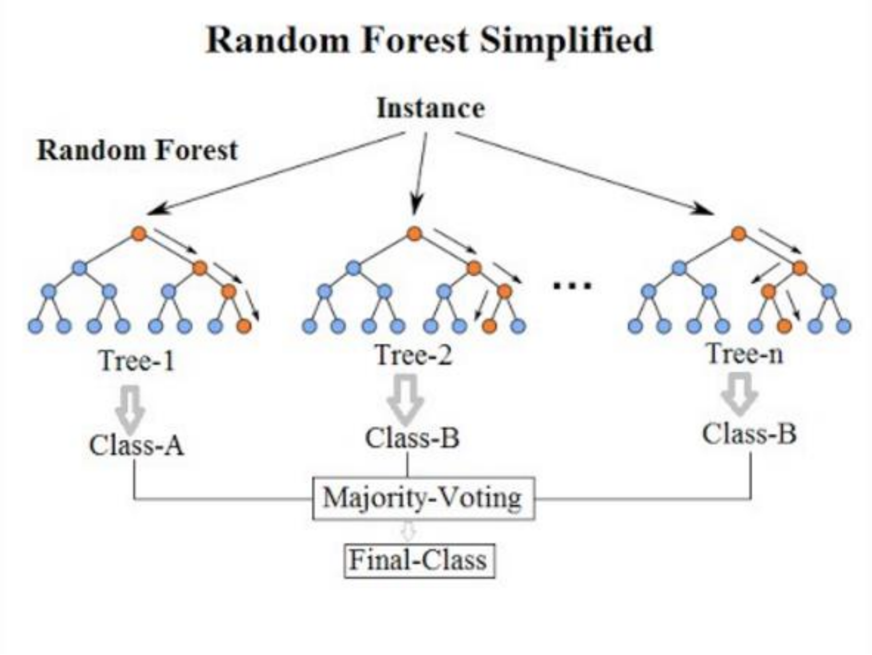


Fig 3.4 Working of Random Forest Algorithm

3.4.3 K Nearest Neighbor

It is a non-parametric algorithm and it assigns labels to new classes based on the majority of classes of its K nearest neighbors in the feature space it is useful when the decision boundary is complex. It can handle multiclass classification problems and is easy to implement.

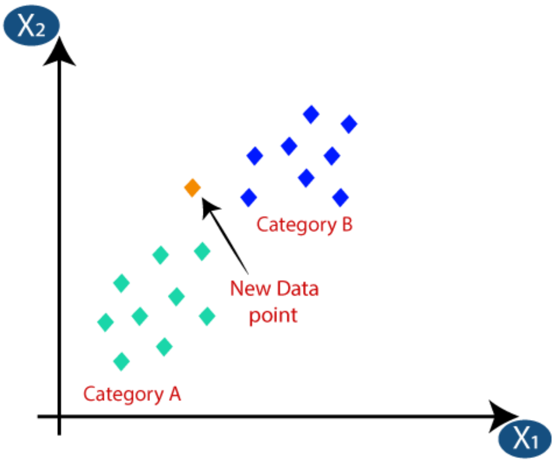


Fig 3.5 Graph of K Nearest Neighbor Algorithm

3.4.4 Support Vector Machines

It can work on both classification and regression. This algorithm is useful for fitting non linear data. It can handle high dimensional data and useful where the dimensions exceed the sample.

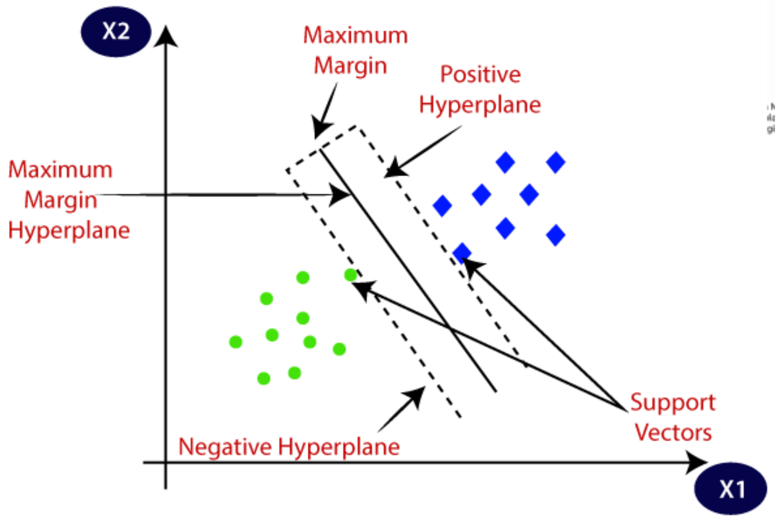


Fig 3.6 Graph of Support Vector Machines Algorithm

3.5 Model Evaluation

Different models were trained with the above algorithms and evaluated on the basis of different performance metrics such as accuracy score, precision, recall, and F1 score. These metrics provide us with data that can be used to derive which algorithm works best with our dataset.

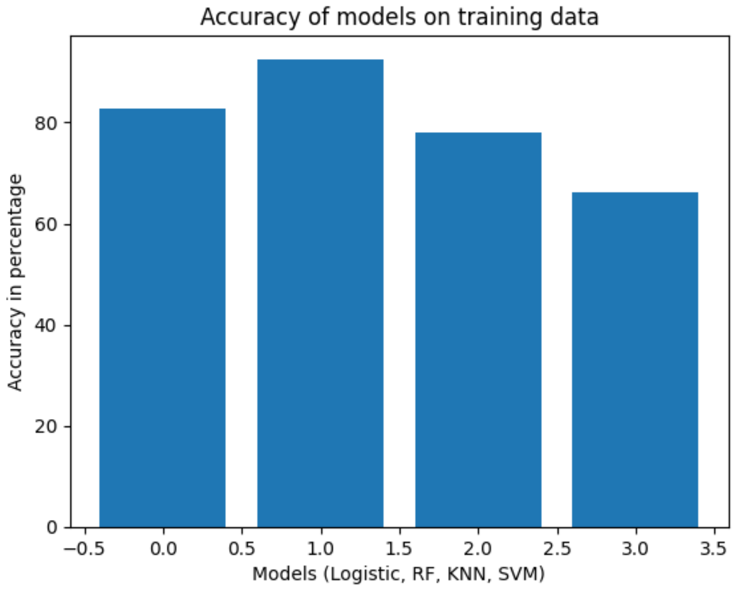


Fig 3.7 Graph showing accuracy of the models (Logistic, RF, KNN, and SVM) on training data

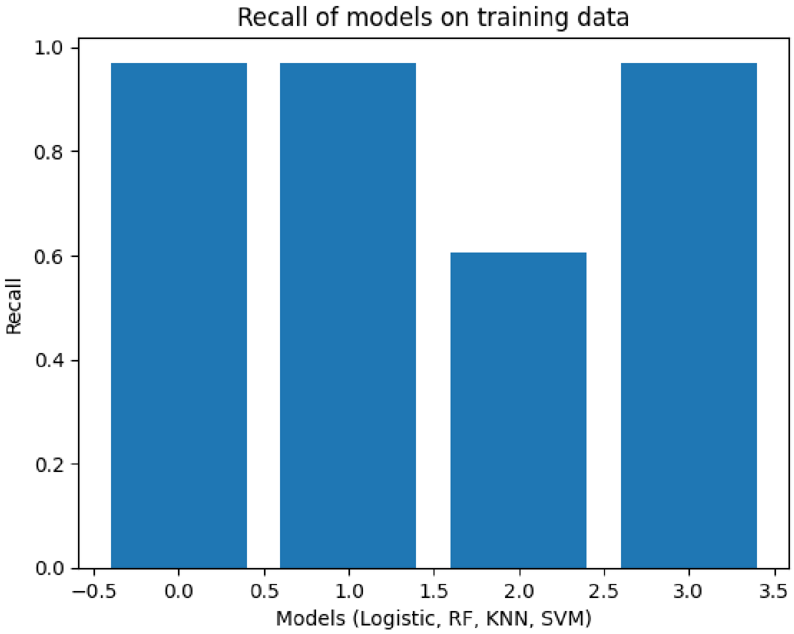


Fig 3.8 Graph showing recall of the models (Logistic, RF, KNN, and SVM) on training data

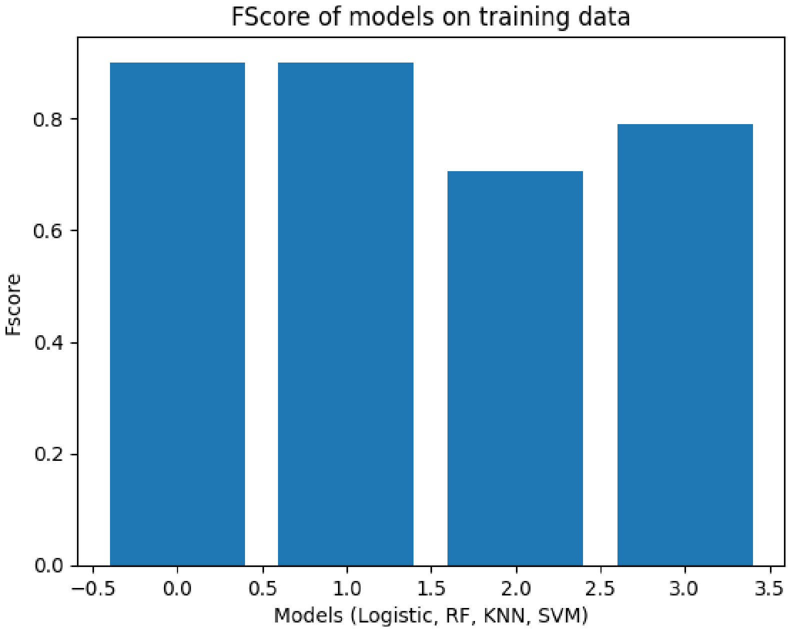


Fig 3.9 Graph showing accuracy of the models (Logistic, RF, KNN, and SVM) on training data

3.6 Experimental Setup

The dataset was split into training and testing in the ratio 80% training: 20% testing. The step was randomized and stratified to provide reproduction of the results.

Chapter 4

Results and Discussion

The performance of Logistic Regression, Random Forest (RF), Support Vector Machines (SVM), and K Nearest Neighbor (KNN) on the Heart Disease Prediction Dataset was analyzed and the results are as follows –

Table 4.1 Showing comparative performance metrices of the tested models

Algorithm	**Accuracy**	**Precision**	**Recall**	**F Score**
Logistic Regression	0.826	0.842	0.970	0.901
Random Forest	0.926	0.842	0.970	0.901
K-Nearest Neighbors	0.781	0.842	0.606	0.705
Support Vector Machines	0.661	0.666	0.970	0.790

4.1 Accuracy

it measures the proportion of correctly classified instances out of the total number of instances in the dataset. It provides an overall assessment of the model's ability to make predictions across all classes. However, using only this metric is not recommended if one of the classes is highly dominant. Here we can see that **Random Forest** algorithm gave the highest accuracy with **92.6%.** Which is because it uses several decision trees to derive the most optimum solution.

4.2 Precision

It measures the proportion of correctly predicted positive values out of the total instances predicted as positive. It focuses on the accuracy of positive predictions Higher precision means the model has lower false positive rates, meaning it makes lesser incorrect positive predictions. Here we can see that the precision of **Logistic Regression, Random Forest** and **KNN** is almost same at **0.842.**

4.3 Recall

It is also called true positive rate or sensitivity it measures the proportion of correctly predicted values out of the total actual positive instances. It focuses on the ability of the model to identify positive instances. High recall means the model has low false negative rates, meaning it can identify most of the positive instances. In the above observation table, we can see that **Logistic, RF, and SVM** have almost identical recall as **0.97.**

4.4 F score

It combines the precision and recall of the model to provide a more balanced evaluation of the models' performance. It denotes the tradeoff between the values of precision and recall, where the higher value of F score denotes a greater balance between the two. Here, **Logistic Regression and Random Forest** are the most balanced with F score of **0.901.**

The above mentioned metrices give us the idea of various machine learning algorithms on our dataset. We can see that various models have various scores but, a higher score doesn't always mean that the algorithms are perfect for our model, we have to fabricate the parameters in relation to our dataset for a better prediction.

The project also has certain limitations such as the dataset available has very low number of tuples (303). Which means that the model might not perform well if the data size is increased. The features provided in the dataset might not be the ones that are really relevant for the prediction of a disease.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

A heart disease prediction model was made using machine learning. For this, various machine learning algorithms were tested for suitability and the most fitting model for the given dataset was used. In this case, that happened to be **Logistic Regression**, as it provides less complex computation and works well on dataset of smaller size. It works with linear dataset.

Random forest had the highest accuracy on the training dataset but that can't be used as the dataset has very low number of entries. Further, Random Forest may overfit the model which is not good, the accuracy of 92% is not realistic as it should be between 70-90%.

Other algorithms such as KNN and SVM performed poorly on this dataset hence they were not used. KNN showed decent accuracy but the recall was extremely low and that's the reason it was not selected. SVM had high recall but the least accuracy hence it wasn't chosen as well.

The major steps involved in making this project were observing trends in the features, training and evaluating the models, comparing the models on the basis of performance metrices and then choosing the most suitable for to make heart disease predictions.

5.2 Future Work

Even though this project performed for the given circumstances, it still has quite a room for improvement in the following aspects:

- Increasing the number of features in the dataset will help in choosing more relevant features for the predictions and thus reduce the amount of noise in the data.
- Increase the number of entries in the dataset hence reduce the chance of underfitting and make a reliable, and much more robust model to predict even a larger set of values with higher accuracy and precision.
- Using Convolutional Neural Networks (CNN) to implement deep learning and analyze time series data and Realtime data to predict heart disease.
- Using external datasets from different sources to train and test the model to make it more general and increase its efficiency for different patient bases.

References

[1] Kaggle. (n.d.). Kaggle Datasets. Retrieved from Heart Disease Dataset | Kaggle

[2] Google Colab. (n.d.). Google Colab. Retrieved from https://colab.research.google.com/

[3] YouTube. (n.d.). YouTube. Retrieved from StatQuest: Logistic Regression - YouTube, Random Forest Algorithm Clearly Explained! - YouTube, Support Vector Machines Part 1 (of 3): Main Ideas!!! - YouTube, K Nearest Neighbors | Intuitive explained | Machine Learning Basics - YouTube

[4] Geeks For Geeks (n.d). GFG retrieved from Random Forest Classifier using Scikit-learn - GeeksforGeeks, Introduction to Support Vector Machines (SVM) - GeeksforGeeks, Logistic Regression in Machine Learning - GeeksforGeeks, K-Nearest Neighbor(KNN) Algorithm - GeeksforGeeks

[5] Stack Overflow. (n.d.). Stack Overflow. Retrieved from https://stackoverflow.com/