A Project

Submitted in partial fulfilment of the completion of the course 19CCE213 Machine Learning and Artificial Intelligence under the Faculty of Computer and Communication Engineering (CCE)

S. No.	Name	Roll No.
1	Byrapu Reddy Sadhana	CB.EN.U4CCE20010
2	Juturu Naga Sai Loknath	CB.EN.U4CCE20025
3	Santosh	CB.EN.U4CCE20053
4	V Srihari Moorthy	CB.EN.U4CCE20060

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE CAMPUS (INDIA)

2021 – 2022

Even Semester

DECLARATION

We hereby declare that the project work entitled, “Heart Disease Prediction Using Machine Learning” submitted to the Department of Computer and Communication Engineering is a record of the original work done by us under the guidance of Ms Suguna G., Faculty, Assistant Professor at Amrita School of Engineering, Amrita Vishwa Vidyapeetham and that it has not been performed for the award of any Degree/Diploma/Associate Fellowship/Fellowship and similar titles if any.

Signature of the Faculty

HEART DISEASE PREDICTION USING MACHINE LEARNING

Suguna G.
Assistant Professor
Amrita School of Engineering
Coimbatore, India
g_suguna@cb.amrita.edu

Byrapu Reddy Sadhana
B.Tech. Computer and Communication Engineering.
Amrita School of Engineering
Coimbatore, India
cb.en.u4cce20010@cb.students.amrita.edu

Juturu Naga Sai Loknath
B.Tech. Computer and Communication Engineering.
Amrita School of Engineering
Coimbatore, India
cb.en.u4cce20025@cb.students.amrita.edu

Santosh
B.Tech. Computer and Communication Engineering.
Amrita School of Engineering
Coimbatore, India
cb.en.u4cce20053@cb.students.amrita.edu

V Srihari Moorthy
B.Tech. Computer and Communication Engineering.
Amrita School of Engineering
Coimbatore, India
cb.en.u4cce20060@cb.students.amrita.edu

Abstract – In most cases, heart disease diagnosis depends on a complex combination of clinical and pathological data. Because of this complexity, there is a significant amount of interest among clinical professionals and researchers regarding efficient and accurate heart disease prediction. In this paper, we develop a heart disease prediction system that can assist medical professionals in predicting heart disease status based on the clinical data of patients. The system will consist of multiple features, including an input clinical data section, ROC curve display section, and prediction performance display section (execute time, accuracy, sensitivity, specificity, and predict result). The paper also discusses the pre-processing methods, classifier performances, and evaluation metrics. We have investigated the accuracy levels of various machine learning techniques such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, and Decision Trees (DT). In the result section, the visualized data shows that the prediction is accurate. The system developed in this study proves to be a novel approach that can be used in the classification of heart disease.

Keywords – Statistical Description and Dispersion, Correlation, Feature Analysis, Classification, K-Nearest Neighbour, Decision Tree, Support Vector Machines, Naïve Bayes

INTRODUCTION

The work proposed in this paper focuses mainly on various data mining practices that are employed in heart disease prediction. The human heart is the principal part of the human body. It regulates blood flow throughout our bodies, any irregularity to the heart can cause distress in other parts of the body. In today's contemporary world, heart disease is one of the primary reasons for the occurrence of most deaths. It may occur due to an unhealthy lifestyle, smoking, alcohol, and high intake of fat which may cause hypertension.[1] According to the World Health Organization, more than 10 million die yearly from heart diseases. A healthy lifestyle and the earliest detection are the only ways to prevent heart-related diseases.

The main challenge in today's healthcare is the provision of quality services and effective accurate diagnoses.[2] Even though heart diseases are found to be more prominent in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in the management of disease lies in the proper time of detection of that disease. The proposed work attempts to detect these heart diseases at an early stage to avoid disastrous consequences. Records of a large set of medical data created by medical experts are available for analyzing and extracting valuable knowledge from it.

Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization.[3] Mostly the medical database consists of discrete information, thereby decision-making becomes a complex task.

A data analysis system that does not handle large amounts of data should be more appropriately categorized as a machine learning system, a statistical data analysis tool, or an experimental system prototype. A system that can only perform data or information retrieval, including finding aggregate values, or that performs deductive query answering in large databases should be more appropriately categorized as a database system, an information retrieval system, or a deductive database system.[3]

In the medical field, machine learning can be used to diagnose, detect, and predict various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease at an early stage.[4] This will help provide effective treatment to patients and avoid severe consequences. This paper presents a details performance analysis using various machine learning techniques, K-means, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, and Decision Tree.[5]

OBJECTIVE

The main objective of doing this research is to present a heart disease prediction model for the prediction of the occurrence of heart disease. Further, this research work is aimed toward identifying the best classification algorithm for identifying the possibility of heart disease in a patient. It is justified by performing a comparative study and analysis using four classification algorithms: Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naïve Bayes and Decision Tree. These are used at different levels of evaluation. Although these are commonly used machine learning algorithms, heart disease prediction is a vital task involving the highest possible accuracy. Hence, the four algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better.

MOTIVATION

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but they are either expensive or are not efficient to calculate the chance of heart disease in humans. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient 24 hours, since it requires more sapience, time and expertise. With a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

LITERATURE SURVEY AND RELATED WORKS

With growing development in the field of medical science alongside machine learning, various experiments and research have been carried out in these recent years releasing relevant significant papers.

- D. Bertsimas, L. Mingardi and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," in

IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3627-3637, Sept. 2021, doi: 10.1109/JBHI.2021.3066347.

- N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.
- Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97 (2021). https://doi.org/10.1007/s12553-020-00505-7
- Dwivedi, A.K. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput & Applic 29, 685–693 (2018). https://doi.org/10.1007/s00521-016-2604-1

METHODOLOGY

1. Existing System

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says "Prevention is better than cure", early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.

2. Proposed System

The proposed work predicts heart disease by exploring the above-mentioned four classification algorithms and carrying out performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into the model which predicts the probability of having heart disease.

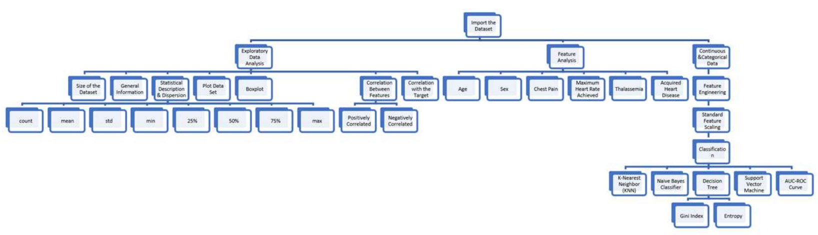

Figure 1 – Generic model to predict heart disease

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 80% of training data is used and 20% of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

3. Attribute Information

The dataset is a combination of 4 different databases, but the primary one is the UCI Cleveland dataset. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features.[6] Therefore, we have used the already processed UCI Cleveland dataset available on the Kaggle website for our analysis.

Serial Number	Attribute	Distinct Values of Attribute
1	Age (in Years)	NIL
2	Sex	Female (0)
		Male (1)
3	Chest Pain	Asymptomatic (0)
		Nonanginal (1)
		Nontypical (2)
		Typical (3)
4	Resting Blood Pressure (mm Hg on admission to the hospital)	NIL
5	Serum Cholesterol Measurement (mg/dl)	NIL
6	Fasting Blood Sugar > 120 mg/dl	False (0)
		True (1)
7	Resting Electrocardiographic Results	Showing probable or definite left ventricular hypertrophy by Estes' criteria (0)
		Normal (1)
		Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) (2)
8	Maximum Heart Rate Achieved	NIL
9	Exercise-Induced Angina	No (0)
		Yes (1)
10	Old Peak – ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)	NIL
11	The slope of the peak exercise ST segment	Down Sloping (1)
		Flat (2)
		Up Sloping (3)
12	Number of major vessels coloured by fluoroscopy	0
		1
		2
		3
13	A blood disorder called thalassemia	Dropped from the dataset previously (0 – NA)
		No blood flow in some parts of the heart (1 – fixed)
		Normal blood flow (2 – normal)
		A blood flow is observed but it is not normal (3 – reversible)
14	Acquired Heart Disease (AHD), Output Class	Normal (0 – No)
		Heart Disease (1 – Yes)

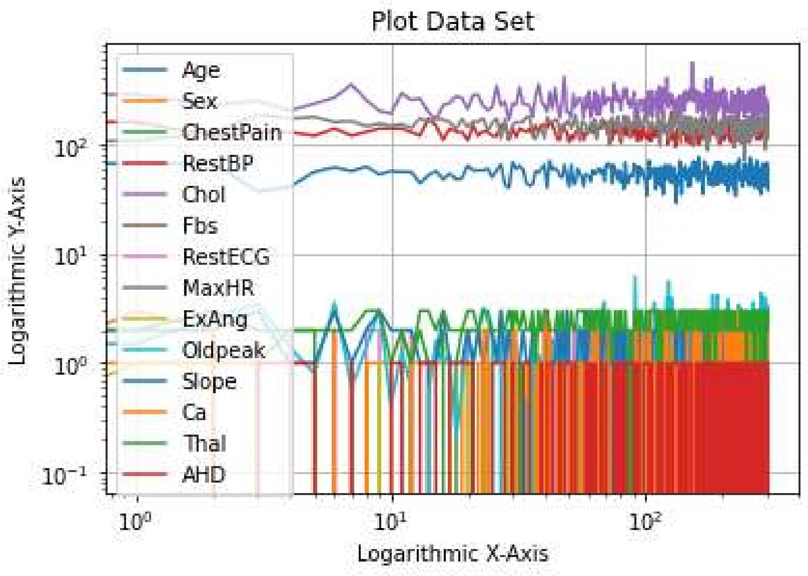

Figure 2 – Plot Data Set

4. Exploratory Data Analysis

```
Size of the dataset -  (303, 14)

General information of the dataset -

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Age        303 non-null    int64
 1   Sex        303 non-null    int64
 2   ChestPain  303 non-null    int64
 3   RestBP     303 non-null    int64
 4   Chol       303 non-null    int64
 5   Fbs        303 non-null    int64
 6   RestECG    303 non-null    int64
 7   MaxHR      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  AHD        303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Figure 3 – General information of the dataset

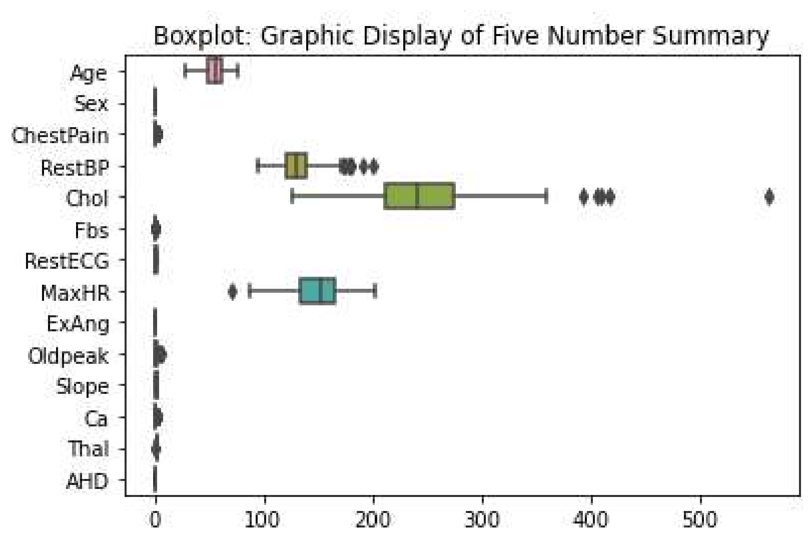

Figure 4 – Boxplot: Graphic Display of Five Number Summary

```
Statistical description and dispersion of the dataset -
              Age         Sex   ChestPain  ...          Ca        Thal         AHD
count  303.000000  303.000000  303.000000  ...  303.000000  303.000000  303.000000
mean    54.438944    0.679868    0.841584  ...    0.663366    2.313531    0.541254
std      9.038662    0.467299    0.960126  ...    0.934375    0.612277    0.499120
min     29.000000    0.000000    0.000000  ...    0.000000    0.000000    0.000000
25%     48.000000    0.000000    0.000000  ...    0.000000    2.000000    0.000000
50%     56.000000    1.000000    1.000000  ...    0.000000    2.000000    1.000000
75%     61.000000    1.000000    1.000000  ...    1.000000    3.000000    1.000000
max     77.000000    1.000000    3.000000  ...    3.000000    3.000000    1.000000

[8 rows x 14 columns]
```

Figure 5 – Statistical Description and Dispersion of the Dataset

It is always better to check the correlation between the features so that we can analyze which feature is negatively correlated and which is positively correlated.

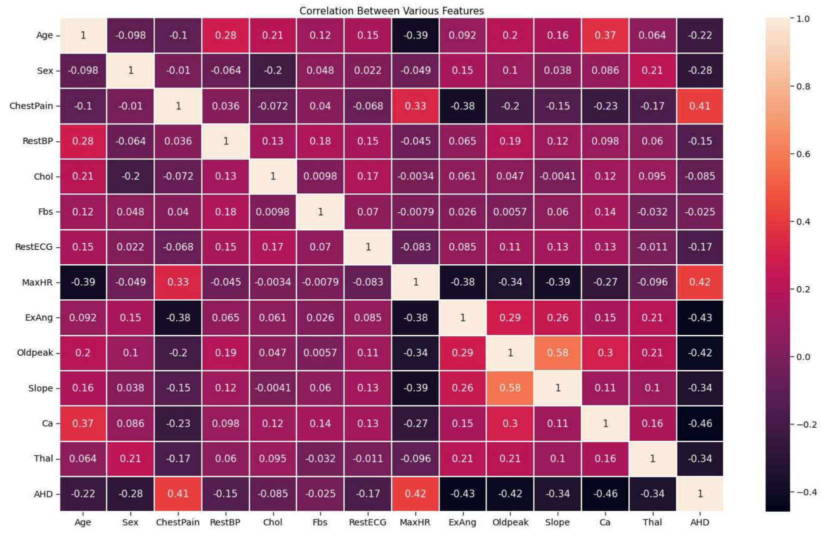

Figure 6 – Correlation Between Various Features

Attribute or feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum, cholesterol, etc. are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

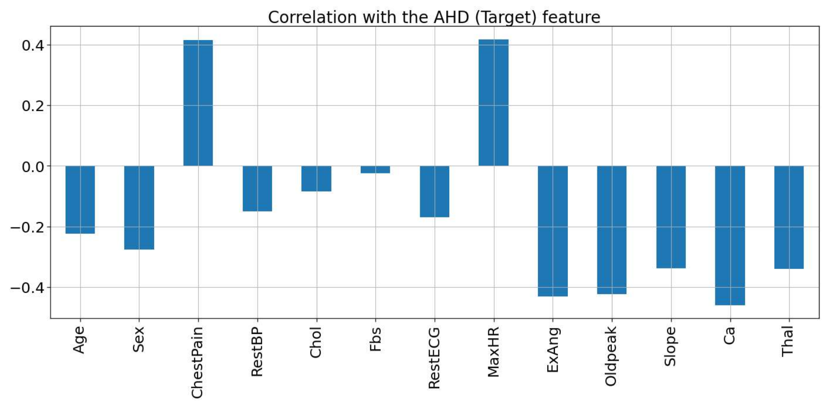

Figure 7 - Correlation with the Acquired Heart Disease (Target) feature

Except for the Chest Pain and Maximum Heart Rate Achieved features, all others are negatively correlated with the AHD (Target) feature.

5. Age Feature Analysis

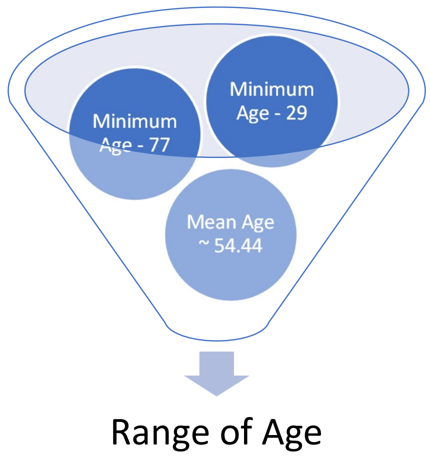

Figure 8 – Range of Age

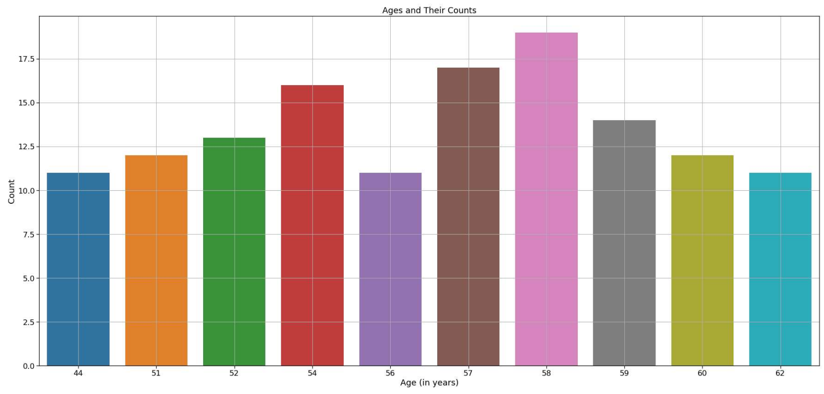

Figure 9 – Ages and Their Counts

We observe that the 58-age group has the highest frequency.

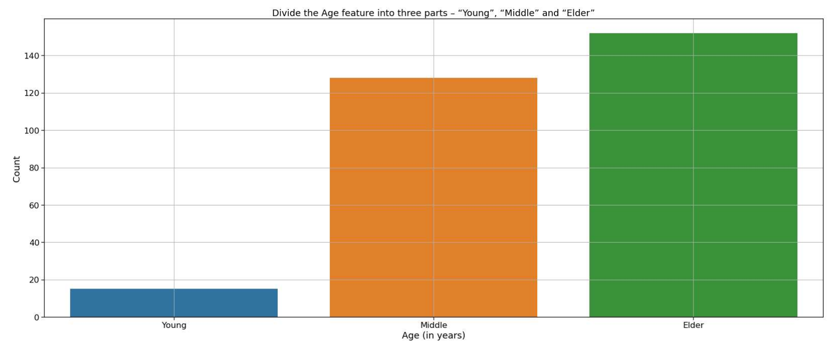

Figure 10 – Divide the Age feature into three parts – Young, Middle, and Elder

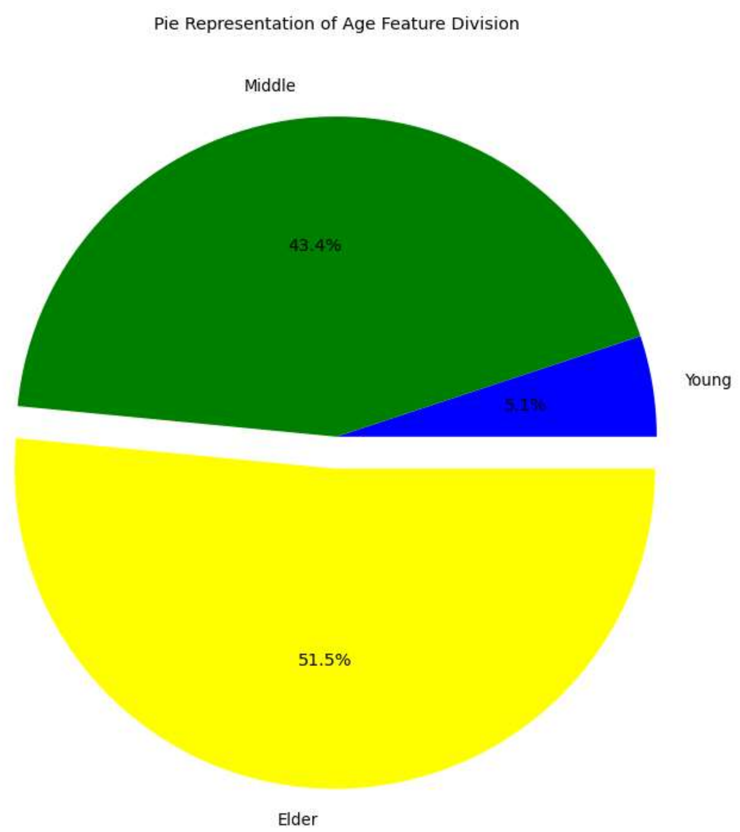

Figure 11 – Pie Representation of Age Feature Division

We observe that elderly people are the most and the young ones are the least affected by heart disease.

6. <u>Sex Feature Analysis</u>

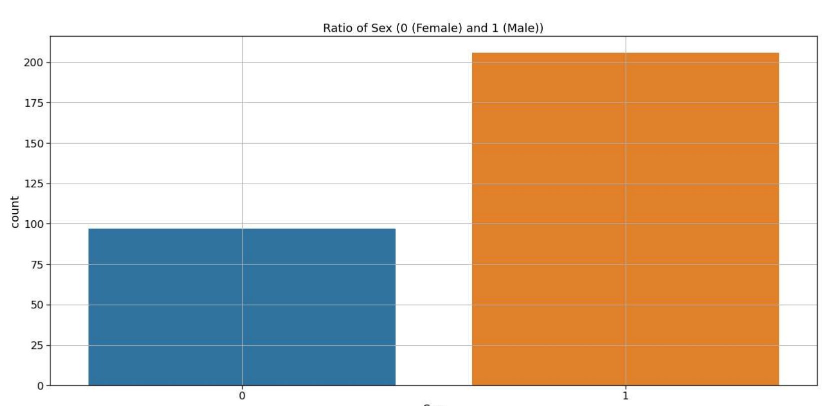

Figure 12 – Ratio of Sex (0 (Female) and 1 (Male)

We observe that the female to male ratio is approximately 1:2.

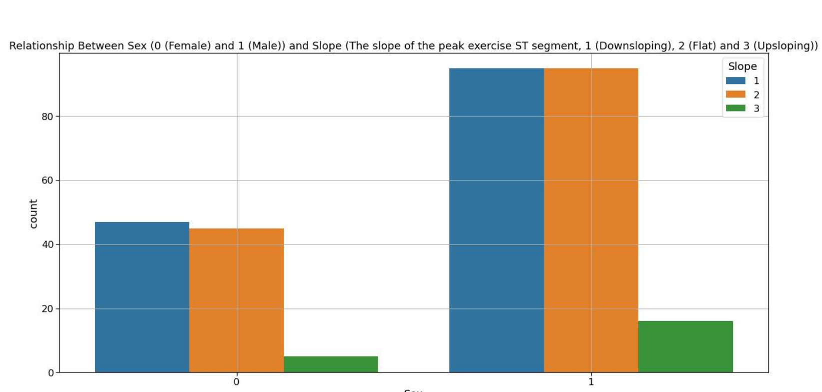

Figure 13 – Relationship Between Sex (0 (Female) and 1 (Male)) and Slope (The slope of the peak exercise ST segment, 1 (Down Sloping), 2 (Flat), and 3 (Up Sloping))

We see that the slope value is higher in the case of males than females.

7. <u>Chest Pain Feature Analysis</u>

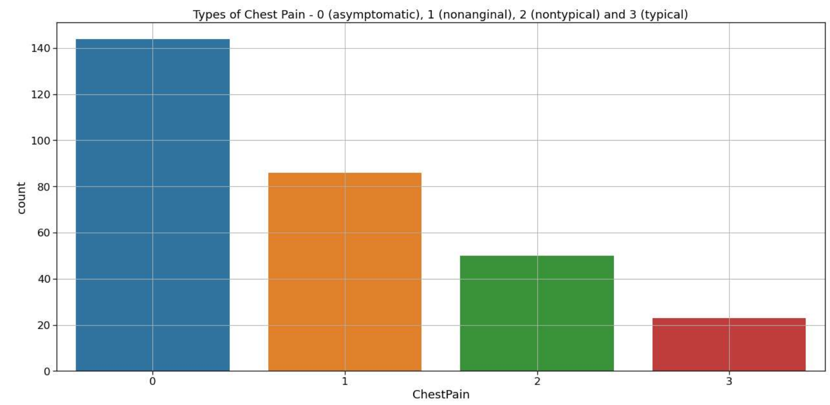

Figure 14 – Types of Chests Pain - 0 (asymptomatic), 1 (nonanginal), 2 (nontypical) and 3 (typical)

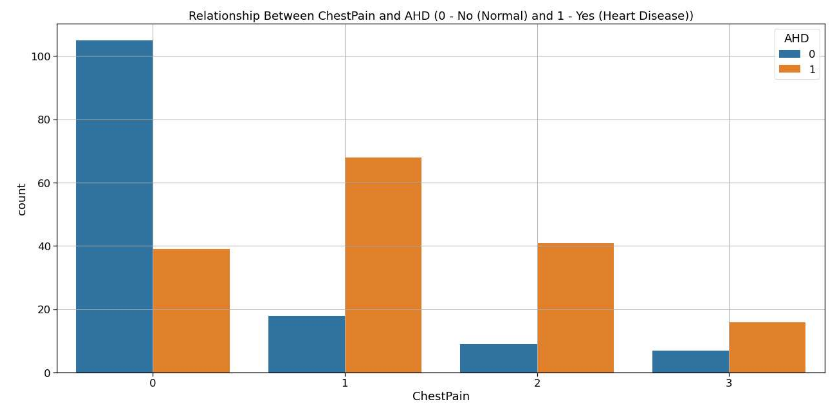

Figure 15 – Relationship Between Chests Pain and AHD (0 - No (Normal) and 1 - Yes (heart disease)

We observe that:

- People having the least chest pain are not likely to have heart disease.
- People having severe chest pain are likely to have heart disease.

8. <u>Maximum Heart Rate Feature Analysis</u>

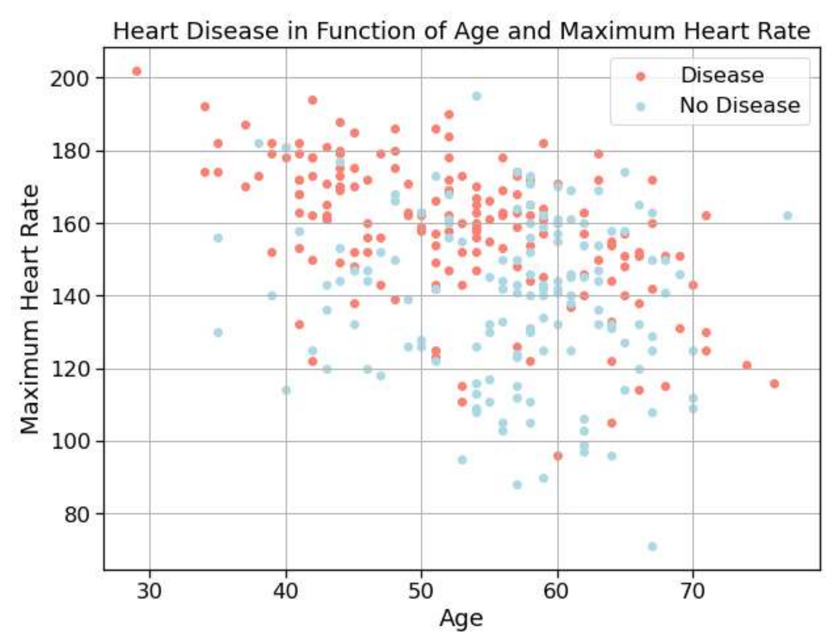

Figure 16 – Heart Disease in Function of Age and Maximum Heart Rate

9. <u>Thalassemia Feature Analysis</u>

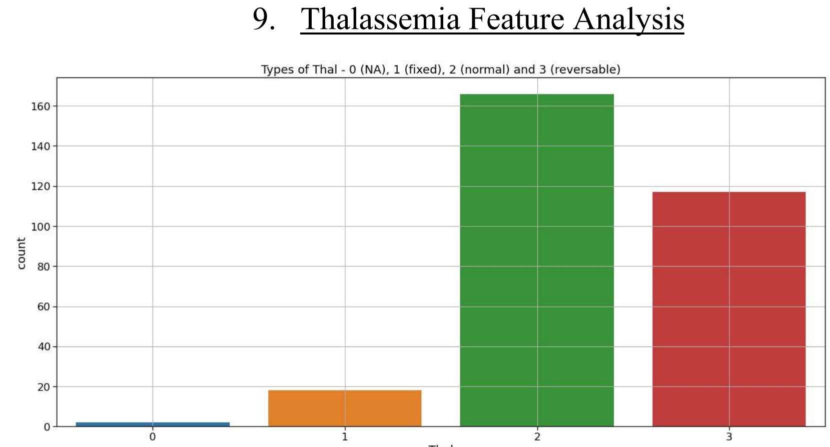

Figure 17 - Types of Thalassemia - 0 (NA), 1 (fixed), 2 (normal) and 3 (reversable)

10. Acquired Heart Disease (AHD) Feature Analysis

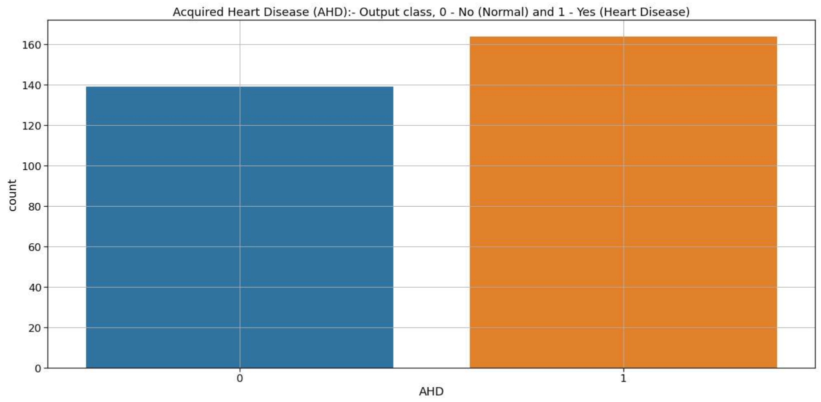

Figure 18 – Acquired Heart Disease (AHD): Output class, 0 - No (Normal) and 1 - Yes (heart disease)

The ratio between 1 and 0 is much less than 1.5 which indicates that the target feature is not imbalanced. So, for a balanced dataset, we can use accuracy scores as evaluation metrics for our model.

11. Complete Description of Continuous and Categorical Data

```
Age : [63 67 37 41 56 62 57 53 44 52 48 54 49 64 58 60 50 66 43 40 69 59 42 55
 61 65 71 51 46 45 39 68 47 34 35 29 70 77 38 74 76]

Sex : [1 0]

ChestPain : [3 0 1 2]

RestBP : [145 160 120 130 140 172 150 110 132 117 135 112 105 124 125 142 128 170
 155 104 180 138 108 134 122 115 118 100 200  94 165 102 152 101 126 174
 148 178 158 192 129 144 123 136 146 106 156 154 114 164]

Chol : [233 286 229 250 204 236 268 354 254 203 192 294 256 263 199 168 239 275
 266 211 283 284 224 206 219 340 226 247 167 230 335 234 177 276 353 243
 225 302 212 330 175 417 197 198 290 253 172 273 213 305 216 304 188 282
 185 232 326 231 269 267 248 360 258 308 245 270 208 264 321 274 325 235
 257 164 141 252 255 201 222 260 182 303 265 309 307 249 186 341 183 407
 217 288 220 209 227 261 174 281 221 205 240 289 318 298 564 246 322 299
 300 293 277 214 207 223 160 394 184 315 409 244 195 196 126 313 259 200
 262 215 228 193 271 210 327 149 295 306 178 237 218 242 319 166 180 311
 278 342 169 187 157 176 241 131]

Fbs : [1 0]

RestECG : [2 0 1]

MaxHR : [150 108 129 187 172 178 160 163 147 155 148 153 142 173 162 174 168 139
 171 144 132 158 114 151 161 179 120 112 137 157 169 165 123 128 152 140
 188 109 125 131 170 113  99 177 141 180 111 143 182 156 115 149 145 146
 175 186 185 159 130 190 136  97 127 154 133 126 202 103 166 164 184 124
 122  96 138  88 105 194 195 106 167  95 192 117 121 116  71 118 181 134
  90]

ExAng : [0 1]

Oldpeak : [2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 3.1 0.4 1.3 0.  0.5 1.6 1.  1.2 0.2 1.8
 3.2 2.4 2.  2.5 2.2 2.8 3.  3.4 6.2 4.  5.6 2.9 0.1 2.1 1.9 4.2 0.9 1.1
 3.8 0.7 0.3 4.4]

Slope : [3 2 1]

Ca : [0 3 2 1]

Thal : [1 2 3 0]

AHD : [1 0]
```

Figure 19 – Part I. Complete Description of Continuous and Categorical Data

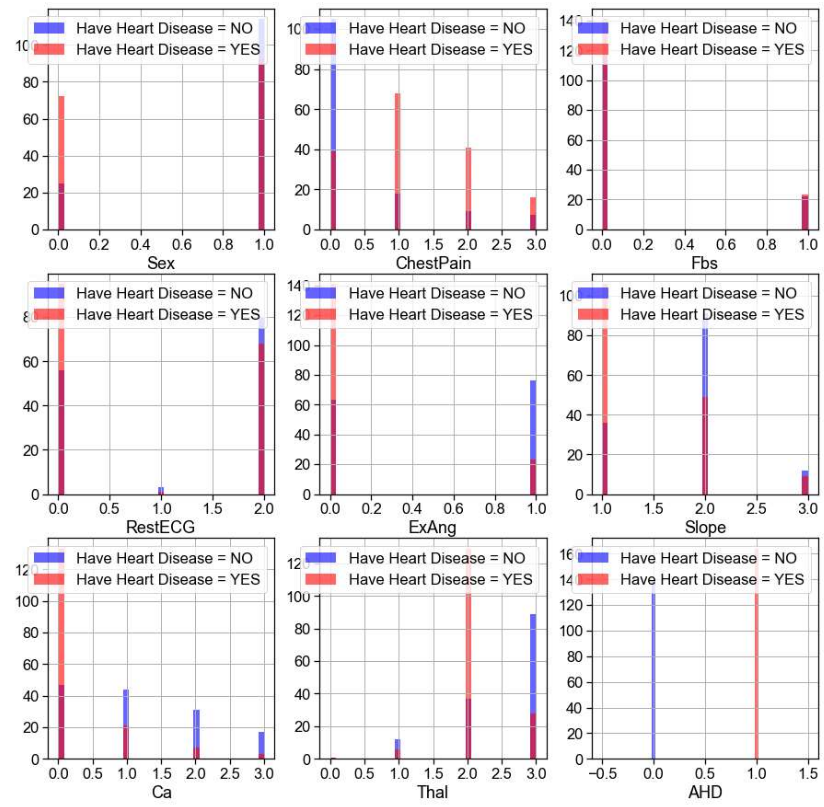

Figure 20 – Part II. Complete Description of Continuous and Categorical Data

- Chest Pain – People with chest pain equal to 1, 2, and 3 are more likely to have heart disease than people with chest pain equal to 0.
- Resting Electrocardiographic Results – People with a value of 0 (showing probable or definite left ventricular hypertrophy by Estes' criteria, which can range from mild symptoms to severe problems) are more likely to have heart disease.
- Exercise-Induced Angina – People with a value of 0 (No) have heart disease more than people with a value of 1 (Yes).
- Slope – People with a slope value equal to 1 (Down-sloping - Signs of Unhealthy Heart) are more likely to have heart disease than people with a slope value equal to 2 (Up-sloping - Better Heart Rate with Exercise) or 3 (Flat - Minimal Change, Typical Healthy Heart).

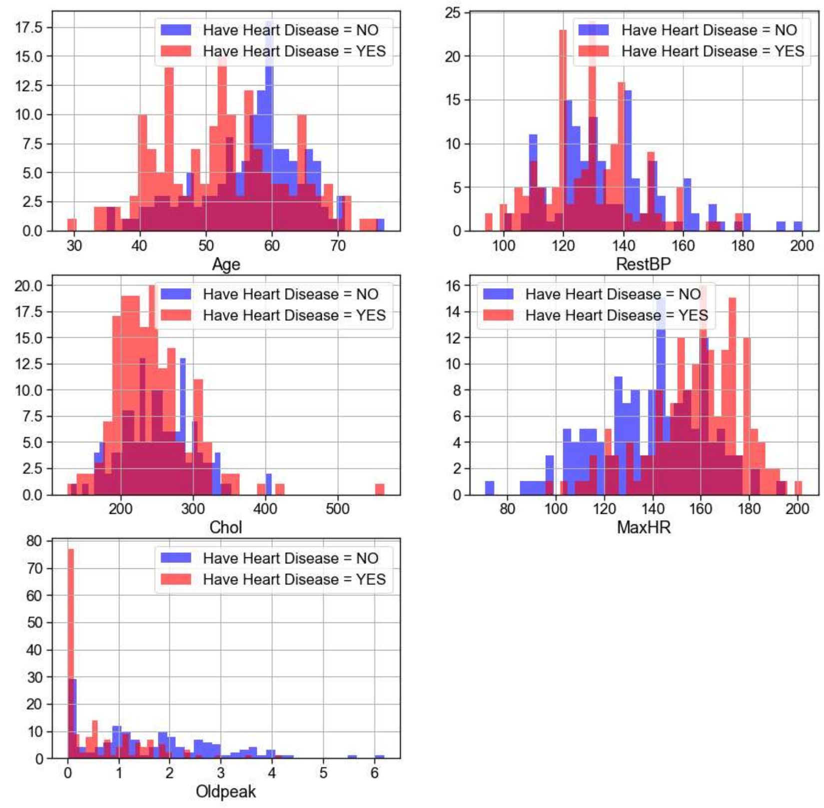

Figure 21 – Part III. Complete Description of Continuous and Categorical Data

- Age – Elderly people (>50 years) are more likely to have heart disease.
- Resting Blood Pressure – Anything between 120-140 (mm Hg on admission to the hospital) is typically a cause for concern.
- Serum Cholesterol Measurement – Anything between 200-300 (mg/dl) is typically a cause for concern.

```
                              Feature Engineering
     Age  RestBP  Chol  MaxHR  Oldpeak  ...  Ca_3  Thal_0  Thal_1  Thal_2  Thal_3
0     63     145   233    150      2.3  ...     0       0       1       0       0
1     67     160   286    108      1.5  ...     1       0       0       1       0
2     67     120   229    129      2.6  ...     0       0       0       0       1
3     37     130   250    187      3.5  ...     0       0       0       1       0
4     41     130   204    172      1.4  ...     0       0       0       1       0
..   ...     ...   ...    ...      ...  ...   ...     ...     ...     ...     ...
298   45     110   264    132      1.2  ...     0       0       0       0       1
299   68     144   193    141      3.4  ...     0       0       0       0       1
300   57     130   131    115      1.2  ...     0       0       0       0       1
301   57     130   236    174      0.0  ...     0       0       0       1       0
302   38     138   175    173      0.0  ...     0       0       0       1       0

[303 rows x 30 columns]
```

Figure 22 – Feature Engineering

Removed the AHD (Target) column from our set of features and categorized all the categorical variables using the get dummies method which will create a separate column for each category.

```
                                        Standard Feature Scaling
          Age    RestBP      Chol     MaxHR  ...  Thal_0  Thal_1  Thal_2  Thal_3
0    0.948726  0.757525 -0.264900  0.017197  ...       0       1       0       0
1    1.392002  1.611220  0.760415 -1.821905  ...       0       0       1       0
2    1.392002 -0.665300 -0.342283 -0.902354  ...       0       0       0       1
3   -1.932564 -0.096170  0.063974  1.637359  ...       0       0       1       0
4   -1.489288 -0.096170 -0.825922  0.980537  ...       0       0       1       0
..        ...       ...       ...       ...  ...     ...     ...     ...     ...
298 -1.046013 -1.234430  0.334813 -0.770990  ...       0       0       0       1
299  1.502821  0.700612 -1.038723 -0.376896  ...       0       0       0       1
300  0.283813 -0.096170 -2.238149 -1.515388  ...       0       0       0       1
301  0.283813 -0.096170 -0.206864  1.068113  ...       0       0       1       0
302 -1.821745  0.359134 -1.386944  1.024325  ...       0       0       1       0

[303 rows x 30 columns]
```

Figure 23 – Standard Feature Scaling

12. Classification

Naive Bayes Classifier

Confusion Matrix		
	28	4
	3	26

Accuracy: 88.52459016

Report

	Precision	Recall	F1-Score	Support
0	0.9	0.88	0.89	32
1	0.87	0.9	0.88	29
Accuracy			0.89	61
Macro Average	0.88	0.89	0.89	61
Weighted Average	0.89	0.89	0.89	61

AUC Score: 91.91810345

Decision Tree Using Gini Index

Confusion Matrix		
	27	5
	27	22

Accuracy: 80.32786885

Report

	Precision	Recall	F1-Score	Support
0	0.79	0.84	0.82	32
1	0.81	0.76	0.79	29
Accuracy			0.8	61
Macro Average	0.8	0.8	0.8	61
Weighted Average	0.8	0.8	0.8	61

AUC Score: 89.92456897

Decision Tree Using Entropy

Confusion Matrix		
	26	6
	3	26

Accuracy: 85.24590164

Report

	Precision	Recall	F1-Score	Support
0	0.9	0.81	0.85	32
1	0.81	0.9	0.85	29
Accuracy			0.85	61
Macro Average	0.85	0.85	0.85	61
Weighted Average	0.86	0.85	0.85	61

AUC Score: 88.41594828

Support Vector Machine (SVM)

Confusion Matrix		
	26	6
	5	24

Accuracy: 81.96721311

Report

	Precision	Recall	F1-Score	Support
0	0.84	0.81	0.83	32
1	0.8	0.83	0.81	29
Accuracy			0.82	61
Macro Average	0.82	0.82	0.82	61
Weighted Average	0.82	0.82	0.82	61

AUC Score: 89.54741379

K-Nearest Neighbour (KNN)

Confusion Matrix		
	28	4
	2	27

Accuracy: 90.16393443

Report

	Precision	Recall	F1-Score	Support
0	0.93	0.88	0.9	32
1	0.87	0.93	0.9	29
Accuracy			0.9	61
Macro Average	0.9	0.9	0.9	61
Weighted Average	0.9	0.9	0.9	61

AUC Score: 93.53448276

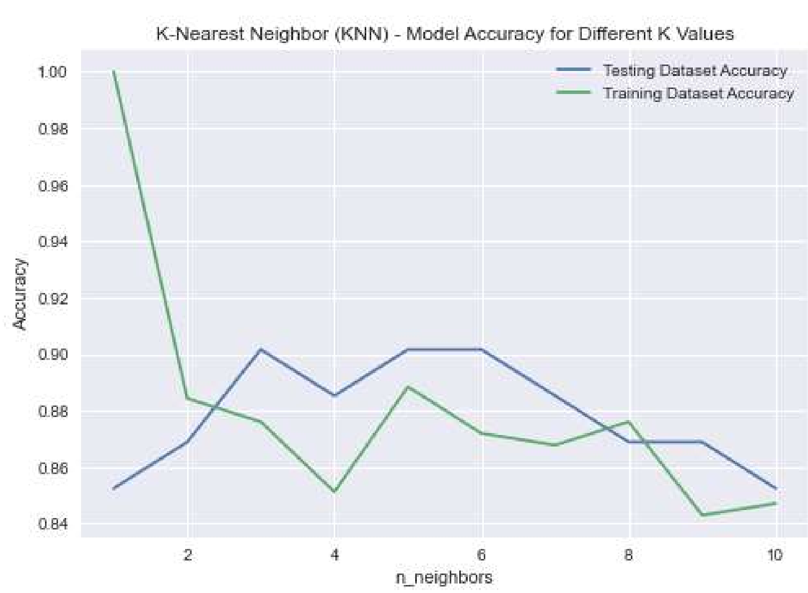

Figure 24 – K-Nearest Neighbour (KNN) - Model Accuracy for Different K Values

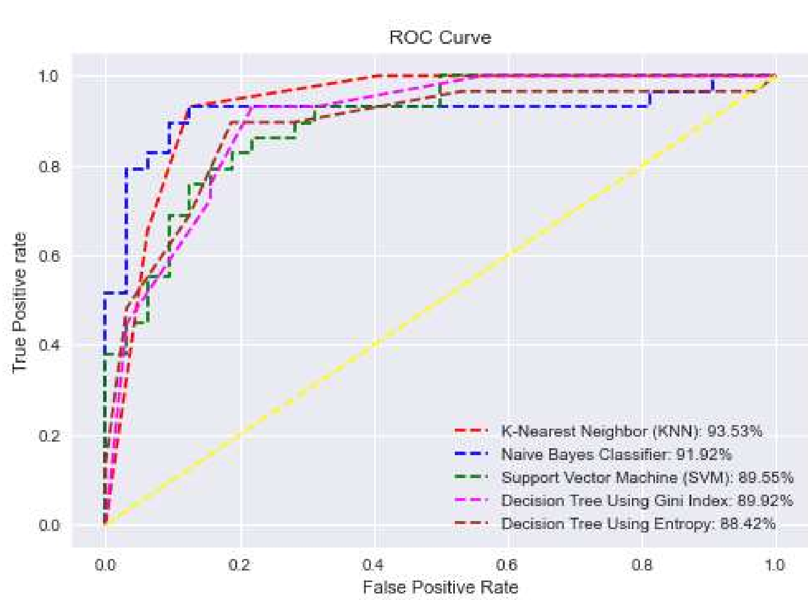

Figure 25 – ROC Curve

CONCLUSION

With the increasing number of deaths due to heart diseases, it has almost become increasingly mandatory to develop a proficient system to predict heart diseases effectively and accurately. This study compares the accuracy score of K-Nearest Neighbour (KNN), Naïve Bayes Classifier, Support Vector Machine, and Decision Tree algorithms for predicting heart disease using the UCI machine learning repository dataset. The result of this study indicates that the K-Nearest Neighbour (KNN) algorithm is the most efficient algorithm with an accuracy score of 90.16% for the prediction of heart disease. In the future, the work can be enhanced by developing a web application based on the K-Nearest Neighbour (KNN) as well as using a larger dataset as compared to the one used in this analysis, which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

ACKNOWLEDGEMENT

I take this opportunity to express a deep sense of gratitude to Dr Sasangan Ramanathan, Dean (School of Engineering), and DDrJayakumar M., Chairperson (Electronics and Communication Engineering), for their cordial support and guidance which helped me in completing the tasks through various stages. It would be my utmost pleasure to express my profound gratitude and deep regards to my course faculty Ms Suguna G. for her exemplary guidance, monitoring, and constant encouragement throughout this project. The blessing, help, and guidance given by her shall carry me a long way in the journey of life on which I'm about to embark. I am obliged to the management of the university for providing quality equipment in the lab, giving me the golden opportunity to carry out this project. Lastly, I thank the almighty, my family, friends, and all others for their constant encouragement.

REFERENCES

[1] T. Nagamani, S. Logeswari, B. Gomathy, Heart Disease Prediction using Data Mining with Mapreduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[2] Avinash Golande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering, Vol 8, pp.944-950, 2019.

[3] Data Mining - Concepts and Techniques, Jiawei Han and Micheline Kamber, Second Edition, 2006

[4] Theresa Princy R, J. Thomas, Human heart Disease Prediction System using Data Mining Techniques, International

Conference on Circuit Power and Computing Technologies, Bangalore, 2016.

[5] Fahd Saleh Alotaibi, Implementation of Machine Learning Model to Predict Heart Failure Disease, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[6] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field, in Machine Learning Paradigms, 2019, pp. 7199.