Heart Disease Prediction

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Candidate Name

Sahitya Akula(AP21110011110)

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

Introduction

Heart disease is a big problem worldwide - indeed, it's a common reason for illness and death. Prompt and exact foresight of heart disease is key for prevention steps and personalized health care. As healthcare data grows at a rapid rate, data mining becomes a valuable tool. It helps extract useful information from complicated datasets. The goal of this project is to utilize data mining techniques. They will be used to craft a strong and effective forecast model for heart disease.
The massive pool of digital medical records, lab results, and patient details existing nowadays offers a unique chance to discover concealed patterns and ties linked to heart disease progression. This project plans to utilize detailed data digging methods to fully explore these datasets. We aim to pinpoint the core risk elements, emerging patterns, and ongoing tendencies that might go unnoticed with conventional approaches.

The fore-seen model we're creating in this project isn't just a tool for recognizing issues early. It also helps formulate tailor-made prevention plans. This project has vast potential. It can leap public health forward by boosting our skill to find people who might get heart disease. This means we can step in on time, leading to better results for patients.

Diving into using data to predict heart disease, we bring together machine learning, choosing important features, and ways to evaluate models. This will be very important. The coming together of actual doctor knowledge and information driven by data suggests a full approach to understanding and lowering the dangers connected with heart disease.

In summary, this project is a forward-thinking move to use the potency of data extraction to tackle the problems heart disease presents. Merging abundant health data with modern analytics methods, we plan to aid in the continuous mission to lessen heart disease's worldwide impact and better the standard of healthcare services.

Project Background

Heart issues, most notably heart disease, remain a big worldwide health problem, causing a lot of sickness and death. Even with progress in medicine, figuring out and handling heart problems early is tricky. Usual testing techniques frequently depend on a small group of factors, which could result in missed issues and late treatment.

1. Complexity of Heart Disease Diagnosis:

Heart disease can appear in different ways. It depends on many things like age, gender, lifestyle, and other health conditions. These many different factors can make it hard for healthcare workers to find one way to diagnose everyone. Seeing this challenge, this project tries to use a whole lot of healthcare data to make a prediction program. This program hopes to diagnose better than the usual methods.

2. Limitations of Conventional Approaches:

RephraseTraditional approaches for diagnosis are priceless. Still, their dependence on hands-on review and a narrow range of factors can be limiting. Such limits can lead to slow discovery and the start of treatment, weakening the power of proactive health actions. This project aims to beat these limits. It uses high-tech data mining methods to find hidden trends and ties within complete health data sets.

3. Role of Data Mining in Healthcare:

Data mining, as a powerful tool in the realm of healthcare analytics, holds the promise of unraveling hidden insights within vast datasets. By systematically analyzing diverse health attributes, this project aims to identify subtle correlations, risk factors, and predictive patterns that may not be immediately evident through conventional means. The utilization of machine learning algorithms enables the creation of predictive models that can enhance the accuracy and timeliness of heart disease prediction.

4. Need for Personalized Healthcare:

The diverse nature of heart disease necessitates a shift toward personalized healthcare solutions. Understanding the unique combination of factors contributing to an individual's cardiovascular health is pivotal for tailoring preventive strategies. The

project responds to the growing demand for more personalized and proactive healthcare approaches by leveraging data mining to create models capable of individualized risk assessment.

5. Importance of Early Detection:

Early detection of heart disease significantly influences the efficacy of interventions and treatment outcomes. By predicting the likelihood of heart disease in its early stages, healthcare professionals can implement targeted preventive measures, lifestyle interventions, and medical treatments, thereby mitigating the progression of the condition and improving long-term prognosis.

6. Overall Project Significance:

The significance of this project lies in its potential to revolutionize the approach to heart disease diagnosis and prevention. By integrating data mining techniques, the project aspires to empower healthcare professionals with a comprehensive tool that not only identifies at-risk individuals but also assists in tailoring interventions based on a nuanced understanding of their health profiles. This, in turn, has the potential to reduce the societal and economic burden of heart disease, enhance the quality of patient care, and contribute to the advancement of personalized medicine in cardiovascular health.

Description of the Project

1. Data Loading and Exploration:

The code begins by importing necessary libraries such as pandas, matplotlib, seaborn, and various machine learning libraries.
The dataset, presumably named "heart.csv," is loaded into a Pandas DataFrame (data).
A correlation matrix is calculated and stored in correlation_matrix.
An "interactions table" is created and populated with correlation coefficients, and both the matrix and table are displayed using heatmap and tabular representations.

2. Data Visualization:

The code includes several data visualizations using seaborn and matplotlib.
It creates a histogram to visualize the distribution of age.
A count plot is used to visualize the distribution of gender (sex).
A bar plot is generated to display the frequency of disease presence based on gender.

3. Outlier Detection:

The code applies Z-score-based outlier detection to identify potential outliers in the dataset.
The threshold for Z-score is set to 3, and the indices and columns of the identified outliers are printed.
A subset of the dataset containing the identified outliers is saved to a CSV file named "outliers.csv."

4. Data Preprocessing:

The outliers are removed from the dataset.
The dataset is split into features (X) and the target variable (Y).
The dataset is further split into training and testing sets using train_test_split.
Standardization is performed using StandardScaler on the features.

5. Model Training and Evaluation:

The code implements several machine learning models for heart disease prediction, including Logistic Regression, Naive Bayes, Random Forest, Decision Tree, and K-Nearest Neighbors.
For each model, it prints the accuracy on both the training and test datasets.
Confusion matrices are generated for each model on the test data.

6. Results Visualization:

The code compiles the results of model accuracy into a DataFrame. A bar plot is created to visually compare the accuracy of different models.

Proposed Solution

The project's primary objective is to develop an effective solution for the early prediction of heart disease, leveraging data mining techniques and machine learning algorithms. The code implements a systematic approach to address the complexity of cardiovascular health assessment, with the following proposed solution components:

1. Comprehensive Data Analysis:

The project begins with a thorough analysis of a dataset containing a variety of health attributes, such as age, gender, blood pressure, cholesterol levels, and more. This comprehensive dataset is a key element in building a predictive model that can encapsulate the intricate nature of heart disease.

2. Correlation Analysis:

The code calculates the correlation matrix, revealing relationships between different health attributes. This step is crucial for identifying potential predictors and understanding how various factors interact, contributing to the overall prediction of heart disease.

3. Outlier Detection and Data Cleaning:

Outliers, which could distort the model's accuracy, are identified and removed from the dataset. This ensures that the predictive model is not adversely influenced by anomalies in the data, contributing to the robustness of the proposed solution.

4. Model Selection and Training:

The code implements a variety of machine learning models, including Logistic Regression, Naive Bayes, Random Forest, Decision Tree, and K-Nearest Neighbors. These models are chosen for their suitability in binary classification tasks, making them well-suited for predicting the presence or absence of heart disease

.

5. Model Evaluation:

The accuracy of each model is assessed both on the training and testing datasets. Confusion matrices are generated, providing a detailed breakdown of true positives, true negatives, false positives, and false negatives. This evaluation is crucial for understanding the model's performance and its ability to generalize to new, unseen data.

6. Comparative Analysis:

The results from each model are compiled into a DataFrame, and a bar plot is created for a visual comparison of their accuracy. This comparative analysis aids in selecting the most effective model for heart disease prediction based on the given dataset.

Model Architecture

Logistic Regression:

Purpose: Logistic Regression is a classic statistical method used for binary classification tasks.
Application: In this project, Logistic Regression serves to model the relationship between the independent variables (health attributes) and the binary outcome of heart disease presence or absence.
Strengths: It provides probabilistic interpretations, is relatively simple to implement, and can handle large datasets efficiently.
Considerations: It assumes a linear relationship between the features and the log-odds of the outcome, which might limit its ability to capture complex interactions.

Decision Tree:

Purpose: Decision Trees recursively split the data based on features to create a tree-like model.
Application: In heart disease prediction, Decision Trees can detect complex patterns and interactions among health attributes.
Strengths: Easy to interpret, can handle both numerical and categorical data, and can capture non-linear relationships.
Considerations: Prone to overfitting, especially when the tree grows too deep without pruning.

Naive Bayes:

Purpose: Naive Bayes is a probabilistic classifier based on Bayes' theorem and the assumption of independence between features.
Application: It's applied here to predict heart disease by calculating the probabilities of a patient having or not having the disease based on the given health attributes.
Strengths: Fast, simple, and effective for text classification and other applications with a large number of features.
Considerations: The assumption of feature independence might not hold in real-world datasets.

Random Forest:

Purpose: Random Forest is an ensemble method that constructs multiple Decision Trees and combines their predictions.
Application: In heart disease prediction, Random Forest combines multiple decision trees to improve accuracy and reduce overfitting.
Strengths: Reduces overfitting, handles large datasets with high dimensionality, and provides feature importance estimation.

Considerations: More complex than a single decision tree, and interpretation might be challenging due to the ensemble nature.

K-Nearest Neighbors (KNN):

Purpose: KNN is a non-parametric, instance-based learning algorithm used for both classification and regression.
Application: In this project, KNN predicts heart disease by finding similar instances (patients) based on their attributes.
Strengths: Simple and easy to understand, especially for smaller datasets, and does not assume any underlying data distribution.
Considerations: Sensitive to irrelevant features and requires careful feature scaling for optimal performance.

Experimentation Details

About Dataset

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

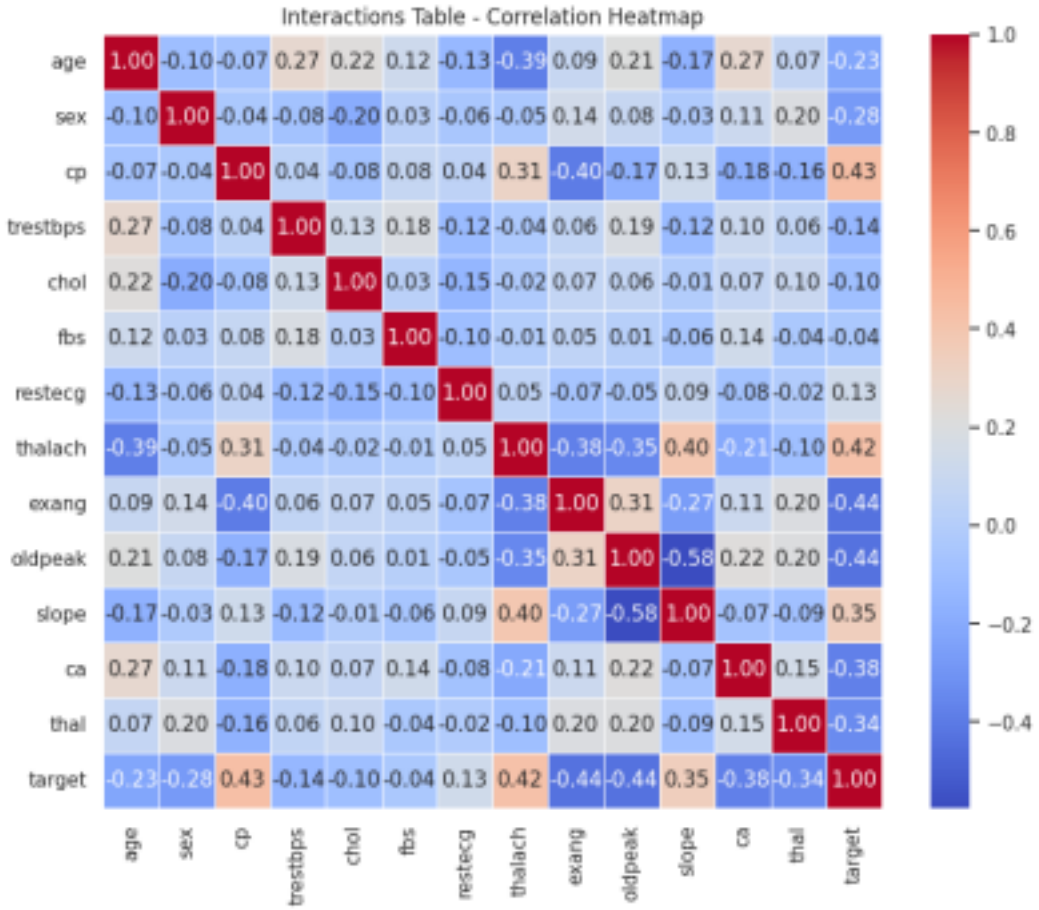
Attribute Information:

- Age
- Sex
- chest pain type (4 values)
- resting blood pressure
- serum cholesterol in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversible defect

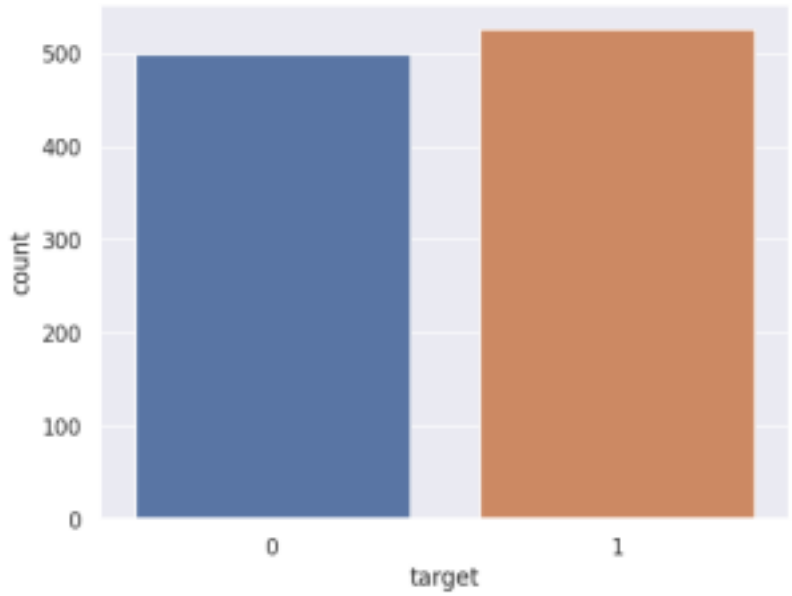
Project Data

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.0	-0.10324	-0.071966	0.271121	0.219823	0.121243	-0.132696	-0.390227	0.088163	0.208137	-0.169105	0.271551	0.072297	-0.229324
sex	-0.10324	1.0	-0.041119	-0.078974	-0.198258	0.0272	-0.055117	-0.049365	0.139157	0.084687	-0.026666	0.111729	0.198424	-0.279501
cp	-0.071966	-0.041119	1.0	0.038177	-0.081641	0.079294	0.043581	0.306839	-0.401513	-0.174733	0.131633	-0.176206	-0.163341	0.434854
trestbps	0.271121	-0.078974	0.038177	1.0	0.127977	0.181767	-0.123794	-0.039264	0.061197	0.187434	-0.120445	0.104554	0.059276	-0.138772
chol	0.219823	-0.198258	-0.081641	0.127977	1.0	0.026917	-0.14741	-0.021772	0.067382	0.06488	-0.014248	0.074259	0.100244	-0.099966
fbs	0.121243	0.0272	0.079294	0.181767	0.026917	1.0	-0.104051	-0.008866	0.049261	0.010859	-0.061902	0.137156	-0.042177	-0.041164
restecg	-0.132696	-0.055117	0.043581	-0.123794	-0.14741	-0.104051	1.0	0.048411	-0.065606	-0.050114	0.086086	-0.078072	-0.020504	0.134468
thalach	-0.390227	-0.049365	0.306839	-0.039264	-0.021772	-0.008866	0.048411	1.0	-0.380281	-0.349796	0.395308	-0.207888	-0.098068	0.422895
exang	0.088163	0.139157	-0.401513	0.061197	0.067382	0.049261	-0.065606	-0.380281	1.0	0.310844	-0.267335	0.107849	0.197201	-0.438029
oldpeak	0.208137	0.084687	-0.174733	0.187434	0.06488	0.010859	-0.050114	-0.349796	0.310844	1.0	-0.575189	0.221816	0.202672	-0.438441
slope	-0.169105	-0.026666	0.131633	-0.120445	-0.014248	-0.061902	0.086086	0.395308	-0.267335	-0.575189	1.0	-0.07344	-0.09409	0.345512
ca	0.271551	0.111729	-0.176206	0.104554	0.074259	0.137156	-0.078072	-0.207888	0.107849	0.221816	-0.07344	1.0	0.149014	-0.382085
thal	0.072297	0.198424	-0.163341	0.059276	0.100244	-0.042177	-0.020504	-0.098068	0.197201	0.202672	-0.09409	0.149014	1.0	-0.337838
target	-0.229324	-0.279501	0.434854	-0.138772	-0.099966	-0.041164	0.134468	0.422895	-0.438029	-0.438441	0.345512	-0.382085	-0.337838	1.0

Correlation matrix is made for the attributes .

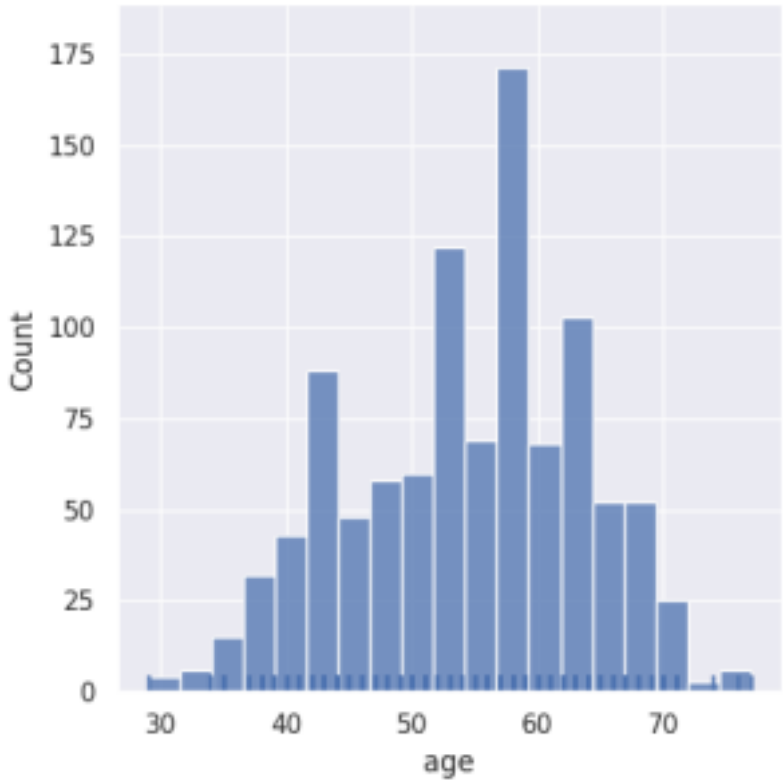


The correlation matrix implemented with heatmap

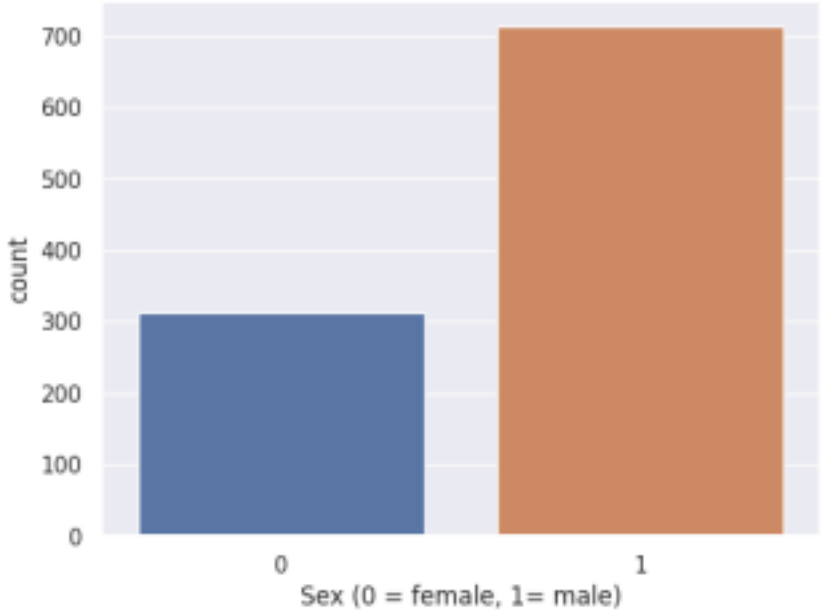


The target attribute shows 1 ---> Defective Heart , 0 ---> Healthy Heart . An extremely

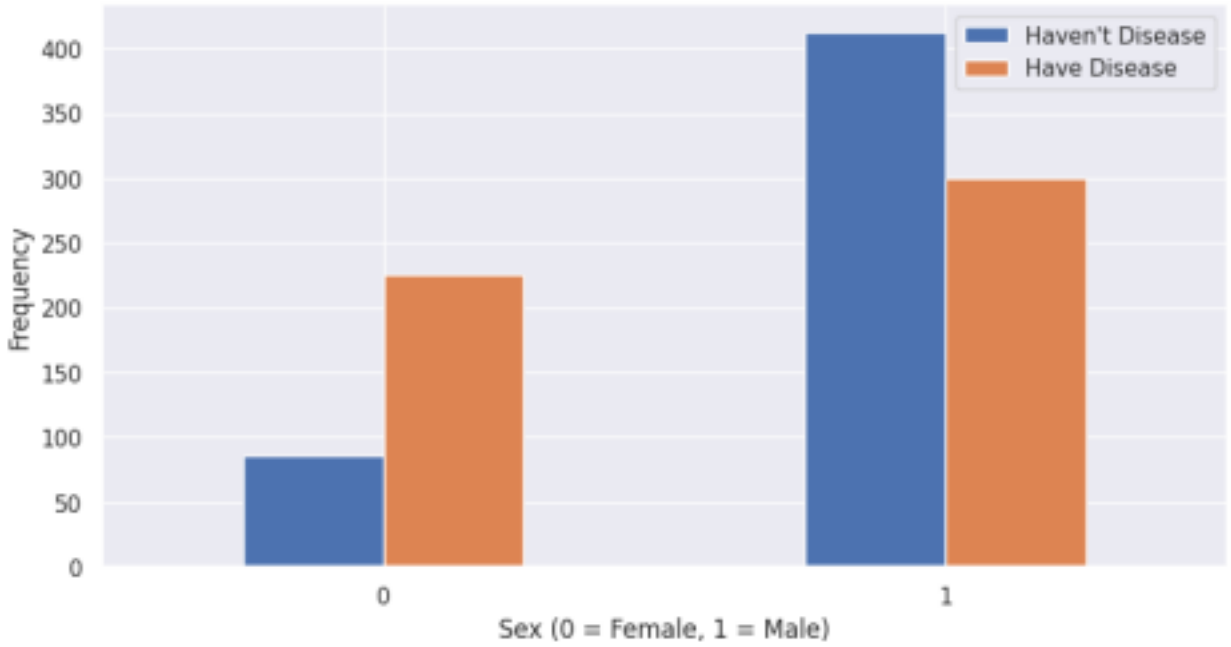
imbalanced dataset can render the whole model training useless and thus, will be of no use.



The Age parameter has mean, standard deviation , minimum and maximum of 54.434146 , 9.072290, 29.00, 77.00 .



Count of male and female in the dataset



Distribution of Heart disease based sex.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	128	204	1	1	156	1	1.0	1	0	0	0
1	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1
2	55	1	0	140	217	0	1	111	1	5.6	0	0	3	0
3	55	1	0	140	217	0	1	111	1	5.6	0	0	3	0
4	62	0	0	160	164	0	0	145	0	6.2	0	3	3	0
5	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1
6	65	0	2	140	417	1	0	157	0	0.8	2	1	2	1
7	52	1	2	138	223	0	1	169	0	0.0	2	4	2	1
8	54	1	1	192	283	0	0	195	0	0.0	2	1	3	0
9	67	0	2	115	564	0	0	160	0	1.6	1	0	3	1

Top 10 , outlier data .An outlier is an observation point that is distant from other observations. The outliers can be a result of a mistake during data collection or it can be just an indication of variance in our data.

```
Accuracy on Training data:  85.85365853658537
Accuracy on Test data:  80.48780487804879
Confussion matrix
[[73 27]
 [13 92]]
```

For Logistic Regression

```
Accuracy on Training data:  100.0
Accuracy on Test data:  100.0
Confusion matrix:
[[100   0]
 [  0 105]]
```

For Decision Tree

```
Accuracy on Training data:  83.90243902439025
Accuracy on Test data:  78.04878048780488
Confussion matrix
[[75 25]
 [20 85]]
```

For Naive Bayes

```
Accuracy on Training data:  85.73170731707317
Accuracy on Test data:  81.46341463414633
Confusion Matrix
[[82 18]
 [20 85]]
```

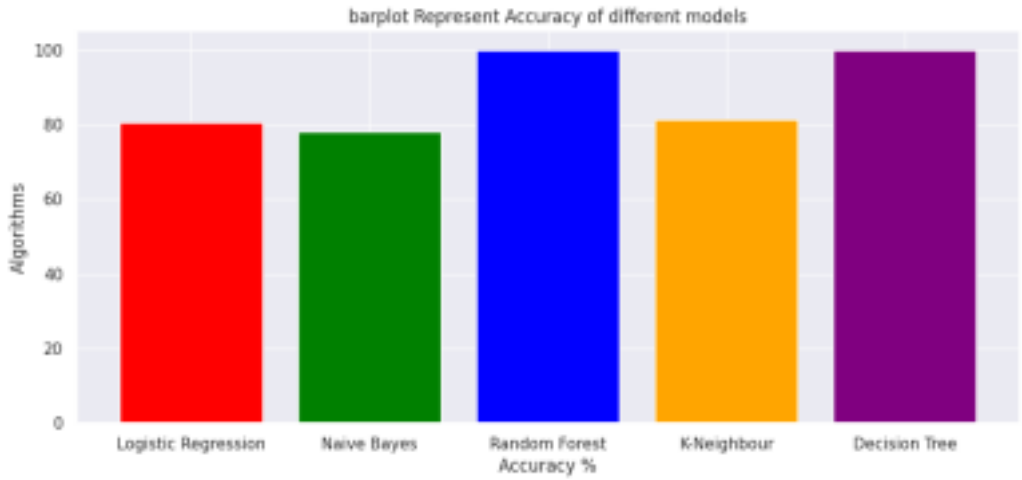
For K-Neighbour


```
Accuracy on Training data:  99.51219512195122
Accuracy on Test data:  100.0
Confussion matrix
[[100   0]
 [  0 105]]
```

For Random Forest

	Models	Accuracy
0	Logistic Regression	80.487805
1	Naive Bayes	78.048780
2	Random Forest	100.000000
3	K-Neighbour	81.463415
4	Decision Tree	100.000000

Results



Results Plot

Conclusion

The correlation analysis shows that a heart diseased patient has a direct relationship with chest-pain(cp) , maximum heart rate achieved(thalch)and the slope of the peak exercise ST segment(slope) .

People mostly of age from 45 to 63 are mostly affected by heart disease. Even people in their late 20’s have shown signs of heart disease.

The dataset is biased towards male , we can infer that males are getting more medical attention than their female counterparts.

Interestingly, more females can be seen with heart disease if we keep consider data of the females than males.

Out of a total of 1025 data entries , only 56 outliers are detected , the number of outliers is relatively less, which helps in the accuracy of further analysis.

We split the data , and train out models like Logistic Regression, Naive Bayers, Random Forest, K Neighbour and Decision Tree. We further check their accuracy . Out of 5 , Decision Tree and Random Forest works better