**Project Report**

Contents

[Introduction 3](#__RefHeading___Toc212_3873381054)

[Part 1: Data Collection, Preprocessing, and EDA 3](#__RefHeading___Toc214_3873381054)

[Independent Features of Dataset: 3](#__RefHeading___Toc216_3873381054)

[Preprocessing Techniques 3](#__RefHeading___Toc218_3873381054)

[Features Engineering 3](#__RefHeading___Toc220_3873381054)

[Visualization Of Data: 4](#__RefHeading___Toc222_3873381054)

[Part 2: Supervised Learning 6](#__RefHeading___Toc226_3873381054)

[1. Regression Models: 6](#__RefHeading___Toc228_3873381054)

[Regression Models Evaluations: 6](#__RefHeading___Toc230_3873381054)

[Comparing Models by Visualization: 6](#__RefHeading___Toc242_3873381054)

[Part 3: Supervised Learning (Classification): 7](#__RefHeading___Toc232_3873381054)

[Accuracy of Classification Models: 7](#__RefHeading___Toc234_3873381054)

[Analysis Of Classification Models: 8](#__RefHeading___Toc236_3873381054)

[Best Models In Classification: 8](#__RefHeading___Toc238_3873381054)

[Conclusion: 8](#__RefHeading___Toc240_3873381054)

[References: 9](#__RefHeading___Toc345_906286157)

Project Report.

# Introduction

we were working on a project in which there is an information about heart disease. We used computer programs to analyze the data and make pictures that help us understand it better. We tried out a few different types of computer methods to guess if someone might have heart disease I have tried both regression and classification model to predict the following person has the disease or not (like a yes or no answer). We did all of this to figure out which computer method works the best for our heart disease info.

# Part 1: Data Collection, Preprocessing, and EDA

### Independent Features of Dataset:

The dataset I chose for this project includes predicting heart disease in patients based on many independent features and one dependent feature (target column) which is output column that our model will predict.

In the data cleaning stage, I used the pandas library to load the csv file into a data frame and checked basic information about the data. There are some int based columns and some are in floating point.

### Preprocessing Techniques

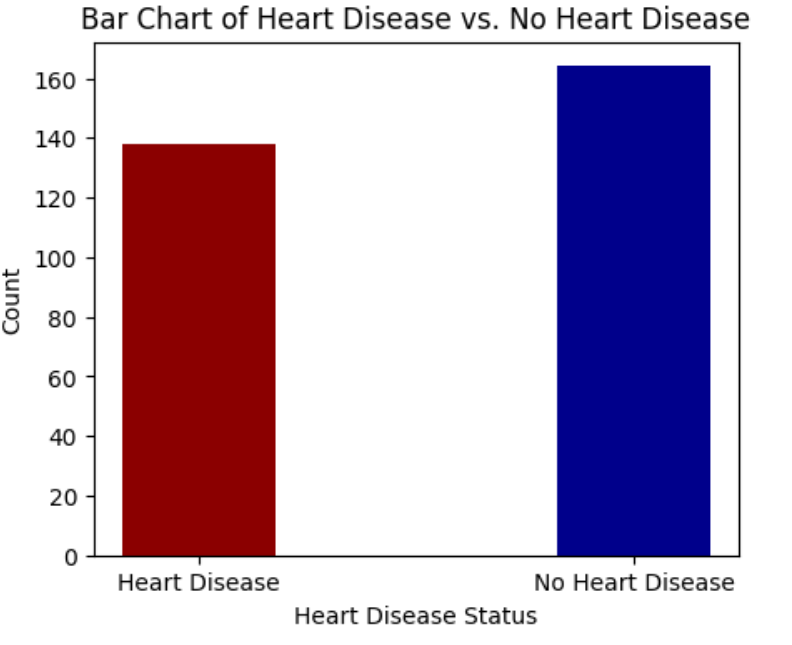
* After reviewing the data I checked the duplication in data. There was some duplicated values in dataset. I found 723 duplicate entries which I removed to keep the data clean.
* There were no missing values in this dataset.
* This is a cleaned version of data set. If we describe the age of data, the minimum age of patient is 29 and Max age of patient is 77. Average age is 54. trestbp(resting bp): average bp = 131.6, min bp = 94, max bp = 200 fbs: average fbs is 0.14 which suggest most patients have fbs < 120 mg/dl (do not have heart disease.)

### Features Engineering

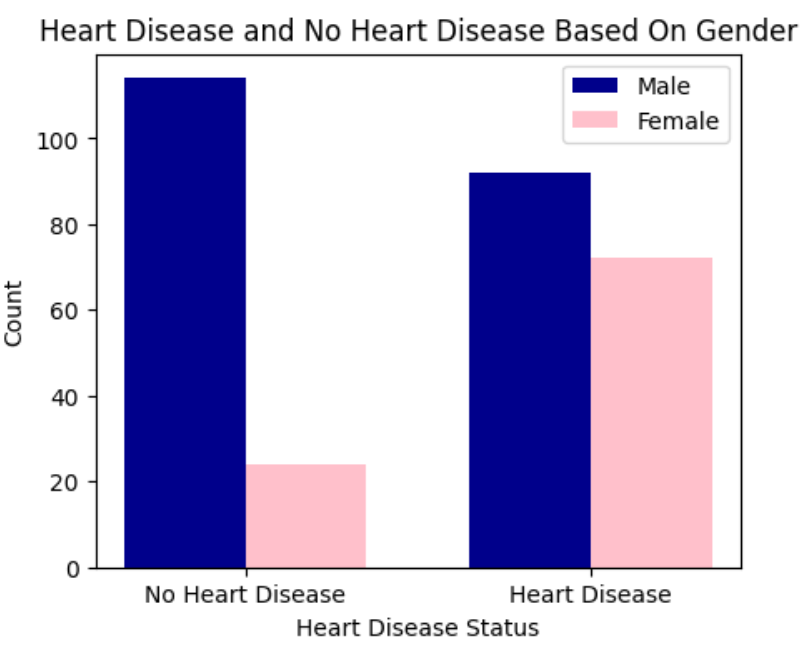
* Percentage of patients with heart disease in age less than 45: 13.57615894039735% Observation: From this data it is observed that the percentage of young patients with age less than 45 who were diagnosed with a heart disease is 13.58%.
* All patients with a normal thalassemia value have reported asymptomatic chest pain. In the given data set all patients who have a normal thalassemia value of 3 have reported only about asymptomatic chest pain. What about having a look at average cholesterol values for both the males and females, which have diabetes, and which do not have diabetes
* Average cholesterol (both males and females): 246.5
* Average cholesterol for males: 239.6
* Average cholesterol for females: 261.3
* Average cholesterol for males with diabetes: 234.12
* Average cholesterol for females with diabetes: 285.83
* Average cholesterol for males with no diabetes: 240.65
* Average cholesterol for females with no diabetes: 257.8

### Visualization Of Data:

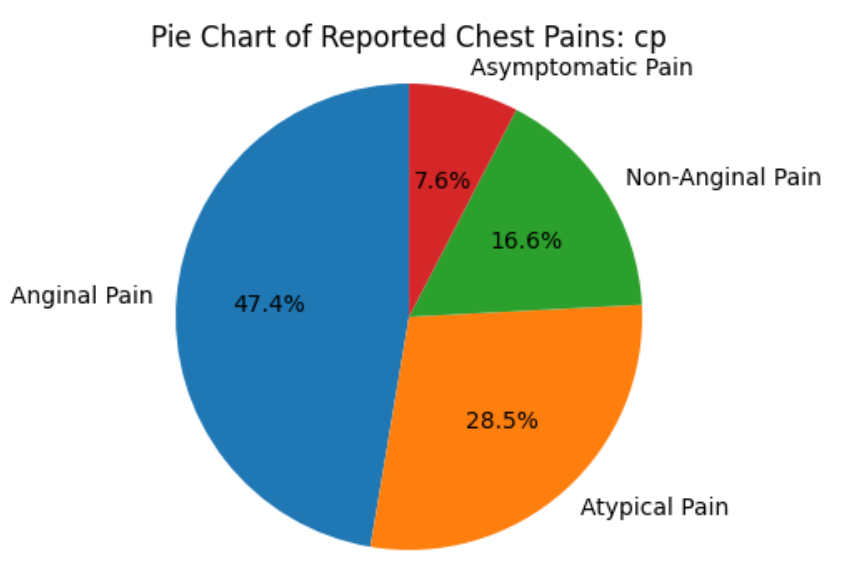
The **Exploratory Data Analysis** involved a variety of visualizations, This is the visualization of people having heart disease and blue hist is showing the people not having the heart disease.



The second Visualization is **heart disease with respect to gender**. How many males and females have this disease. Male average is more than females with respect to heart disease and not having heart disease.



#### Pie Chart for Chest pain:

People have reported Anginal Pain, 28% people have reported Atypical Pains, 16% of people have people have reported Anginal Pain, 28% people have reported Atypical Pains, 16% of people have reported Non-Anginal Pain, and the least percentage of reported pain is Asymptomatic Pain with 7%reported Non-Anginal Pain, and the least percentage of reported pain is Asymptomatic Pain with 7%

* The presence of diabetes does not consistently coincide with the presence of a heart disease, as observed in the bar chart above, patients with diabetes will not necessarily have heart disease.
* The cholesterol levels show a consistent range across both heart disease patients and non-heart disease patients

## Part 2: Supervised Learning

### 1. Regression Models:

Now I have prepared the data for Machine learning to train the model. According to the following I have applied these Regression models on this dataset.

#### 2. Support Vector Regressor:

I have applied Support vector regressor model on this model with these hyper parameters kernel='sigmoid', C=1.0, epsilon=0.2 initialized with a “sigmoid” kernel and an epsilon of “0.2.” This model give the **accuracy of 54%.**

#### Random Forest Regressor

I have applied Random Forest Regressor with these parameters max\_depth=10, n\_estimators=300, random\_state=0 and this model has given the **accuracy of 84%**

## Regression Models Evaluations:

To evaluate the performance of these models, we used three widely recognized metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The results showed a significant difference in the accuracy of predictions between the two models.

**The random Forest regressor**

* Mean squared error of 0.29
* Mean squared error of 0.13
* Root mean squared error of 0.37

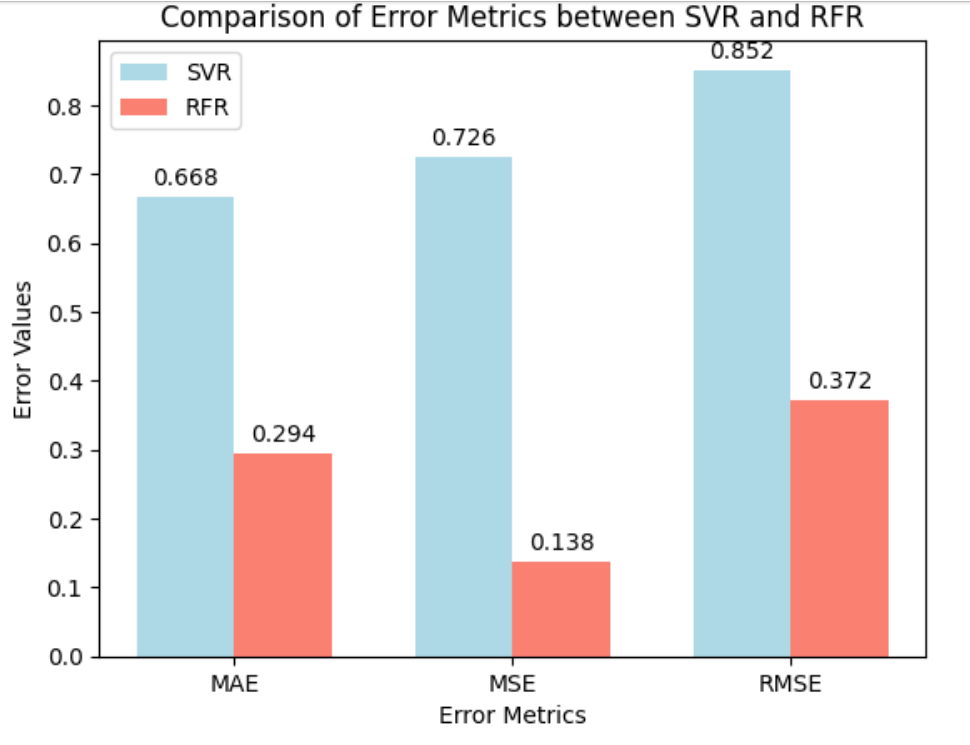
**In support vector regressor**

* Mean Absolute Error of 0.66,
* Mean square error of 0.72,
* Root mean square error of 0.85

## Comparing Models by Visualization:

I compared the performance of these models by using graph. I plotted a graph between Mean Absolute Error ,Mean Squared Error and Root mean squared Error.

However, on the different run, SVR showed some weird results too, which were handled while experimentation.



## Part 3: Supervised Learning (Classification):

Supervised learning classification models to predict diabetes.

**Logistic Regression:**

I have applied Logistic Regression with these hyper parameters max\_iter=300, random\_state=0, solver='lib linear' and it id giving the accuracy of **85.25%**

**Random Forest Classifier**

I applied Random Forest Classifier with these hyper parameters max\_depth=7, n\_estimators=300 and accuracy is **83%**

**Support Vector Machine:**

I have applied SVM by using these hyper parameter gamma='auto', kernel='linear' and accuracy is **83.61%.**

**XGBoost Classifier:**

XGB also perform very good but the accuracy is 81.97%.

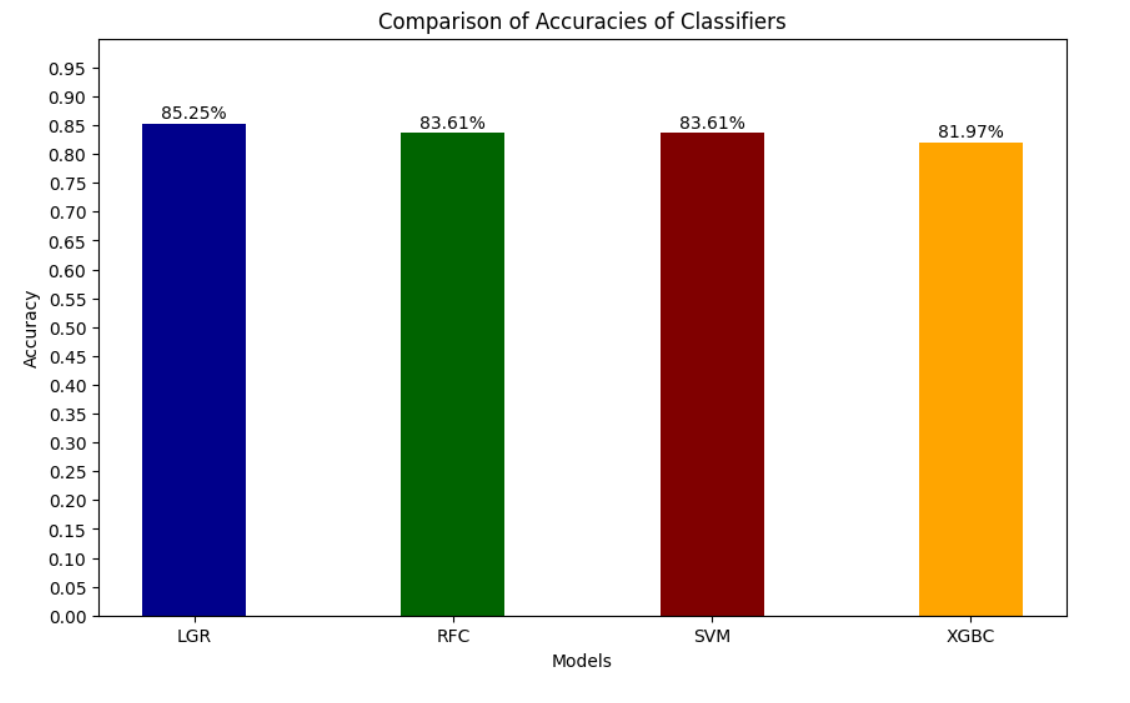
## Accuracy of Classification Models:

The Logistic Regression model achieves an accuracy of 85.25%. On the other hand, the Random Forest Classifier exhibited 83%. The accuracy of Support vector machine is 83.61 % . **In this Classification model Logistic Regression and Random Forest Regressor are performing the best.**

## Analysis Of Classification Models:

Now its time to summarize the project, we did a detailed analysis of predicting heart disease among patients, keeping in view key features and attributes, through pre-processing data, performing necessary EDA to find relevancy and important insights about the data, regression and classification machine learning’s models.

the performance of different models and methods. Key insights emerged during the process, identifying hypertension and glucose\_level with highest feature importance in our dataset.



## Best Models In Classification:

In Classification model Logistic Regression perform the best with accuracy is 85.25%. And in regression model Random Forest Classifier has 84% Accuracy.

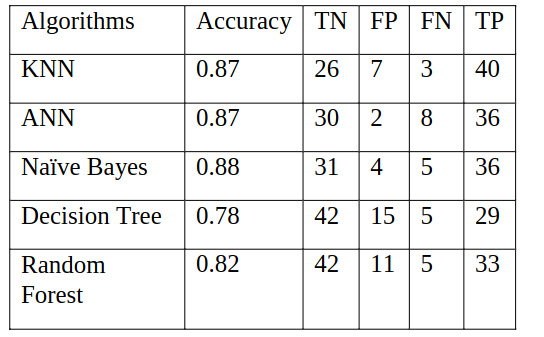
## Comparison With Research Paper:

[1] A. J. Kharya, "Heart Disease Prediction," ResearchGate, 2021: https://www.researchgate.net/publication/349140147\_Heart\_Disease\_Prediction

I have compare my dataset with Heart Disease Prediction, which is solve with different

techniques. They have used different machine and deep neural models to predict the heart disease.

* Artificial Neural Networks.
* K-Nearest Neighbors.
* Naive Bayes.
* Decision Tree
* Random Forest

We have used following Regressions and Classification Models.

* Support Vector Regressor
* Random Forest Regressor
* Random Forest Classifier
* Logistic Regression
* Support vector Machine Classifier
* XGBoost

|  |  |
| --- | --- |
| **Models** | **Accuracy** |
| Support Vector Regressor | 54% |
| Random Forest Regressor | 84% |
| Random Forest Classifier | 83.61 |
| Logistic Regression | 85.25 |
| XGBoost | 81.97% |
| Support vector Machine Classifier | 83.61% |

After observation I came to see there is a common model Random Forest Classifier. Their accuracy is 82% and mine accuracy is 84%. my model is performing little better. The structure of our Random Forest Classifier is RandomForestClassifier (max\_depth=7, n\_estimators=300).

If we want to improve little more there are many options to go into some complex neural network. We can go for neural networks. It will be very beneficial for accuracy and improvement.

## Conclusion:

Our study shows how different factors can affect the prediction of heart disease. It is important to choose the right model for analysis and keep improving our methods. This will help us get better and more useful results in healthcare analytics. If we do better Exploratory Data analyses than our model will perform best. Otherwise model would get confused in outliers.

## References:

DataCamp (2023). "Learn Data Science Online," in DataCamp .Available: [https://www.datacamp.com](https://www.datacamp.com/).

ExcelR (2023). "Data Science Course," in ExcelR. Available: <https://excelr.in/data-science-course>.

[1] A. J. Kharya, "Heart Disease Prediction," ResearchGate, 2021: <https://www.researchgate.net/publication/349140147_Heart_Disease_Prediction>