**HEART FAILURE PREDICTION**

**[DAB 304-006]**

**A picture containing hydrozoan

Description automatically generated**

**GROUP MEMBERS:**

**SURBHI PATEL**

**RICHA PATEL**

**DHARMIKKUMAR PATEL**

**RAJ BHALODWALA**

**JOYAL PATEL**

**TABLE OF CONTENTS**

CHAPTER PAGE

INTRODUCTION 1

BUSINESS OBJECTIVES 1

METHODOLOGY

3.1 – DATA COLLECTION 2

3.2 – DATA CLEANING 3

3.3 – EXPLORATORY DATA ANALYSIS 3

3.4 – DATA VISUALIZATION 5

3.5 – BUILD AND APPLY SUITABLE ALGORITHM 7

3.5.1 LOGISTIC REGRESSION………………………………………………………………………………………………………. 8

3.5.2 NAÏVE BAYES MODEL………………………………………………………………………………………………………… 9

3.5.3 KNN MODEL……………………………………………………………………………………………………………………. 10

3.5.4 SUPPORT VECTOR MACHINE MODEL………………………………………………………………………………. 11

3.5.5 DECESION TREE………………………………………………………………………………………………………………. 12

3.5.6 RANDOM FOREST MODEL………………………………………………………………………………………………. 13

CONCLUSION 13

CONTRIBUTION 14

REFERENCES 14

1. **Introduction:**

According to estimates, 17.9 million people (about the population of New York) die from cardiovascular diseases (CVDs) each year, which accounts for 31% of all fatalities worldwide. Heart attacks and strokes account for four out of every five CVD fatalities, with premature deaths accounting for one-third of these deaths among those under the age of 70. This dataset comprises 11 variables that can be used to predict a potential heart illness. Heart failure is a common occurrence brought on by CVDs. Heart disease is divided into three categories: coronary heart disease, cardiomyopathy, and cardiovascular disease. The phrase "heart disease" comprises several disorders that affect the heart and blood arteries and how the fluid gets into the bloodstream and circulates there in the body. Cardiovascular disease (CVD) is a leading cause of mortality and a variety of illnesses. The diagnosis of a disease is a crucial and challenging task in medicine.

The automation of this work is particularly beneficial since medical diagnosis is regarded as a critical yet challenging task that must be completed swiftly and effectively. Unfortunately, despite the lack of resources in some areas, not all doctors are experts in every field. Data mining may be used to identify hidden patterns and information that may contribute to good decision making. This plays a significant role for healthcare professionals in making accurate judgments and offering quality services to the public. A machine learning model can be very helpful in the early diagnosis and care of people with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors including hypertension, diabetes, hyperlipidaemia, or previously existing illness)

1. **Business Objectives:**

The major aims of establishing this project are:

1. To use Logistic Regression to create a machine learning model that can forecast the likelihood of developing heart disease in the future.
2. To detect considerable risk based on a medical dataset which may lead to heart disease.
3. To study feature selection algorithms and comprehend their working concept.
4. **Methodology:**

* Data Collection
* Data Cleaning (MS Excel, Python)
* Exploratory Data Analysis (Python)
* Data Visualization (Python)
* Build and apply suitable algorithms (Python)

**3.1 Data Collection:**

The dataset is publicly available on the Kaggle Website. It provides patient information which includes over 918 records and 12 attributes. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

Dataset link: [Heart Failure Prediction Dataset | Kaggle](https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)

**Attribute Information:**

1. **Age**: age of the patient [years]
2. **Sex**: sex of the patient [M: Male, F: Female]
3. **ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. **RestingBP**: resting blood pressure [mm Hg]
5. **Cholesterol**: serum cholesterol [mm/dl]
6. **FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. **RestingECG**: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
9. **ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]
10. **Oldpeak**: oldpeak = ST [Numeric value measured in depression]
11. **ST\_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. **HeartDisease**: output class [1: heart disease, 0: Normal]
    1. **Data Cleaning:**

**Table

Description automatically generated**

From the above, it can be clearly seen that the dataset is pretty clean to use for visualization as well as build a model as there are no null values in the dataset and also there is relatively strong datatype as per the column mentioned. There is nothing to do with the cleaning part other than finding null values.

* 1. **Exploratory Data Analysis:**

**Graphical user interface, text, application, email

Description automatically generated**

First, for EDA, the main important part is to work with categorical column. Above are some codes based on finding the unique value for categorical columns. After finding unique value, label encoding method is used for converting categorical columns to numerical. It is just because of creating multiple models and for that numerical value should have in the dataset.

A picture containing graphical user interface

Description automatically generated

The concept of correlation describes the connections between one or more variables. These factors might be characteristics of the raw data that were utilised to forecast our target variable. A statistical method known as correlation shows how one variable moves or changes in connection to another one. It provides us with a general understanding of how closely the two variables are related.

**Graphical user interface, application

Description automatically generated**

It shows that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive.

* 1. **Data Visualization:**

**A screenshot of a computer

Description automatically generated with medium confidence**

The ratio of having heart problem in a person is almost equal to the person not having a heart disease. While comparing for male and female the ratio of having heart disease is much more in Male that is 400 counts higher than having in Female.

Chart, treemap chart

Description automatically generated

For the chest pain patient with ASY type pain has the highest ratio of heart disease with 400 counts whereas for other types of the numbers of patient are below 100. TA has almost balanced ratio of heart patient.

Chart

Description automatically generated

Measuring the heart disease for RestingECG of normal patient has equal chance of having heart disease and for the rest of the ratio is almost same but less than normal patient.

A screenshot of a computer

Description automatically generated with medium confidence

The patient who has ExerciseAngina has more heart problems and if they don’t have Exercise Angina then the ratio of not having heart disease is very more. So ExerciseAngina is directly proportional to heart rate.

A screenshot of a computer

Description automatically generated with medium confidence

For ST\_Slope if the slope is up there are very low chance of having heart problems whereas for flat slope it is opposite. Whereas for Down ST\_Slow both disease and no-disease have very less ratio compared to up and Flat slop.

* 1. **Build and apply suitable algorithms:**
     1. **Logistic Regression**

**Graphical user interface, text, application, email

Description automatically generated**

Chart, treemap chart, square

Description automatically generated

This matrix shows that it has 65 true negative rate, 17 false positive, 12 false negative while true positive cases are 90.

**3.5.2 Naïve Bayes Model:**

Graphical user interface, text, application, email

Description automatically generated

**Chart, treemap chart

Description automatically generated**

This matrix shows that it has 63 true negative rate, 19 false positive, 14 false negative while true positive cases are 88.

**3.5.3 KNN Model:**

**Graphical user interface, text, application

Description automatically generated**

**Chart, treemap chart

Description automatically generated**

This matrix shows that it has 57 true negative rate, 25 false positive, 34 false negative while true positive cases are 68.

* + 1. **Support vector machine model:**

**Graphical user interface, text, application, email

Description automatically generated**

**Chart, treemap chart

Description automatically generated**

This matrix shows that it has 54 true negative rate, 28 false positive, 27 false negative while true positive cases are 75.

* + 1. **Decision Tree:**

**Text

Description automatically generated**

**Chart, treemap chart

Description automatically generated**

This matrix shows that it has 62 true negative rate, 20 false positive, 22 false negative while true positive cases are 80.

**3.5.6 Random forest Model:**

Graphical user interface, text, application, email

Description automatically generated

Chart, treemap chart

Description automatically generated

This matrix shows that it has 62 true negative rate, 20 false positive, 11 false negative while true positive cases are 91.

**Summary:**

Table

Description automatically generated

* From the table of the accuracy for different model, it can be said that logistic regression gives a better accuracy as compared to another model. But the highest recall score is 0.892 for the random forest model.
* Overall, we can say that logistic regression gives a better result.

1. **Conclusion**

The application of data mining techniques in healthcare, namely in the early diagnosis of heart disease, was the major emphasis of our work. Heart disease is a serious condition that can result in death. Data mining techniques were applied utilising the following algorithm, KNN, , Decision Tree, Naive Bayes, logistic regression and Random Forest. We tested performance based on Accuracy, TN, FP, FN and TP rate and in some algorithm.

Overall, we can conclude from the given data that the ratio of the heart diseases is higher in Male that is 55%. For the chest pain patient with ASY type pain has the highest ratio of heart disease with 400 counts. Measuring the heart disease of normal patient has equal chance of having heart disease and for the rest the ratio is 1:10. For ST\_Slope if the slope is up there are very low chance of having heart problems whereas for flat slope it is opposite. The patient who has ExerciseAngina has more heart problems and if they don’t have Exercise Angina then the ratio of not having heart disease is very more. So ExerciseAngina is directly proportional to heart rate. It shows that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive. After performing Accuracy on every model, we can conclude that the Logistic Regression gives the best result out of all the model for the data accuracy test.

**Contribution:**

Surbhi Patel: All 6 models

Richa Patel: Introduction part, data collection

Dharmikkumar Patel: Data Visualization

Raj Bhalodwala: Page designing and index

Joyal Patel: Data cleaning, EDA

**References:**

* <https://www.kaggle.com/code/temitayobadewole/heart-diseases-analysis-and-prediction>
* <https://plotly.com/python/templates/>