BMS COLLEGE OF ENGINEERING

(Autonomous College under VTU)

Bull Temple Road, Basavanagudi, Bangalore – 560019

An PROJECT report on

"Heart-Disease prediction"

Submitted in partial fulfillment of the requirements for the award of degree

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

By

CHIRANJEEVI NAYAK B , (1BM20IS402)

SIDDHARTH MANOJ , (1BM19IS157)

SUSHMITHA R ,(1BM20IS412)

Under the guidance of

SHOBANA T S, Assistant Professor

Department of Information Science and Engineering

2021-22

BMS COLLEGE OF ENGINEERING

(Autonomous College under VTU)
Bull Temple Road, Basavanagudi, Bangalore – 560019

Department of Information Science and Engineering

CERTIFICATE

This is to certify that the project entitled "**Heart-disease prediction**" is a bona-fide work carried out by CHIRANJEEVI NAYAK B **(1BM20IS402), SIDDHARTH MANOJ (1BM19IS157) & SUSHMITHA R(1BM20IS412)** in partial fulfillment for the award of degree of Bachelor of Engineering in **Information Science and Engineering** from **Visvesvaraya Technological University, Belgaum** during the year **2021-2022**. It is certified that all corrections/suggestions indicated for Internal Assessments have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

Signature of the Faculty
Name and Designation

Signature of the HOD
Name and Designation

Table of Contents

Page No.

Abstract……4

Introduction…….……..5

Problem Statement……6

Literature Survey……...6

System Requirement Specifications …….8

System Design / flow diagram …….9

Implementation…..……...10

i. Logistic regression……10

ii. Random Forest……..……11

Test Results ……...…….12

Conclusion……...14

Refrences ……...…….15

ABSTRACT

In recent times, Heart Disease prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This paper makes use of heart disease dataset available in UCI machine learning repository. The proposed work predicts the chances of Heart Disease and classifies patient's risk level by implementing different data mining techniques such as Logistic Regression and Random Forest.

INTRODUCTION

The work proposed in this paper focus mainly on various data mining practices that are employed in heart disease prediction. Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of body. Any sort of disturbance to normal functioning of the heart can be classified as a Heart disease. In today's contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension . According to the World Health Organization more than 10 million die due to Heart diseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases.

The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage . This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyse the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This paper presents performance analysis of various ML techniques such as Logistic Regression for predicting heart disease at an early stage.

PROBLEM STATEMENT:

In this era of stress and anxiety, chronic diseases are a normal occurrence. The only way to control mortality rate is through early prediction and proper treatment of these. So, we are trying for an early prediction and analysis of heart disease by considering multiple various parameters like age, gender, blood pressure, heart rate, diabetes, and more. Since numerous factors are involved, the prediction is a bit challenging.

LITERATURE SURVEY:

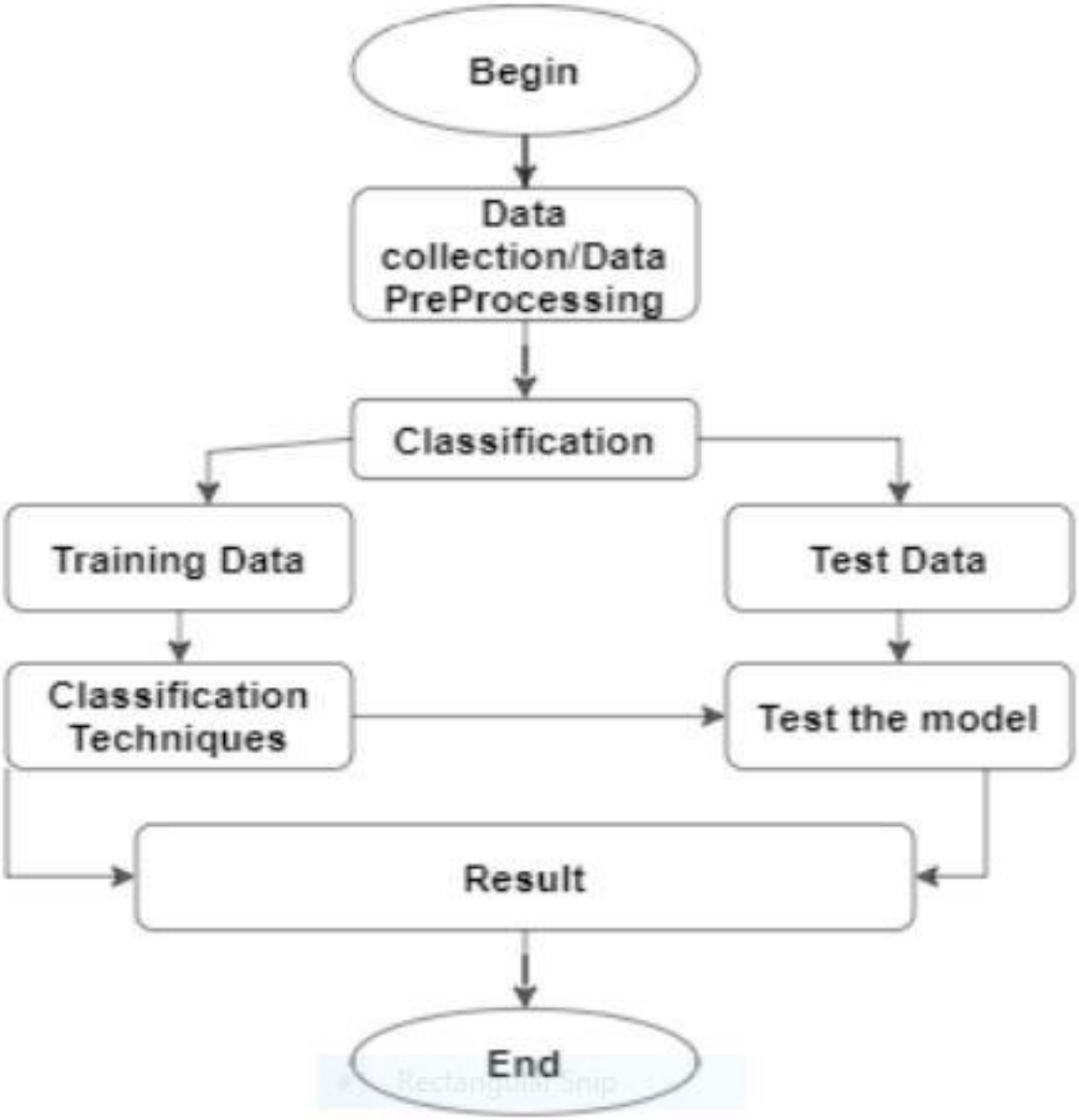
Sr. No	Title -Authors	Year – Publication Site	Description
1.	"Cardiovascular Disease Prediction Using Machine Learning Models" - Atharv Nikam, Sanket Bhandari, Shamla Mantri.	Dec-2022, PuneCon	7 classifiers were used. Decision tree classifier generated the highest accuracy of 73.12%. It's found that BMI plays an essential role. (11 Attributes) But there was a difference in testing and training accuracy due to smaller and hard data in the data set.
2.	"Heart Disease Prediction Using Machine Learning Techniques" - Galla Bindhika, Munaga Meghana, Manchuri Sathvika Reddy.	April – 2020, IRJET	The results showed a good accuracy standard. By introducing Random Forest classification, we can find the prediction rate without equipment. (11 Attributes) But the prediction results is not accurate. Data mining techniques does not help to provide effective decision making. It cannot handle enormous datasets.

3.	"Early and Accurate Prediction of Heart Disease Using Machine Learning Model" - Sairabi H. Mujawar	Oct-2018, IJIRCCE	By using 2 different algorithms, the one with the 93%for presence of HD. 89%for absence of HD accuracy was determined. (14 attributes) While it doesn't work well with large data and accuracy is low, it also follows the black box nature so it is computationally expensive to work on.

SYSTEM REQUIREMENTS AND SPECIFICATIONS

- Windows or Linux operating system.
- Processor: Intel or high
- RAM: 1024 MB
- Space on disk: minimum 100m
- dataset about the heart disease prediction.
- python version 3 or above
- IDE(jupyter notebook) or GOOGLE COLAB

System Design / flow diagram:



IMPLEMENTATION

Logistic regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predictingthe categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shapedlogistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on itsweight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous anddiscrete datasets.

Random Forest

- Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.
- A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.
- The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions.

Test_Results

Logistic Regression

```
[40] model = LogisticRegression()
```

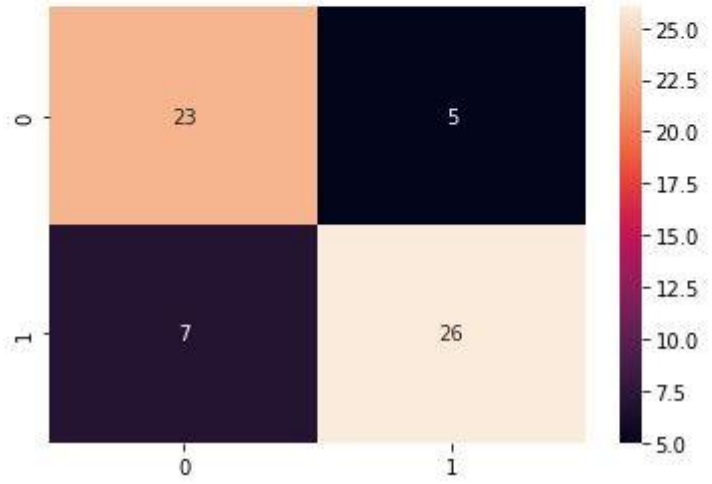
```
# training the LogisticRegression model with Training data
model.fit(X_train, Y_train)

# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

print('Accuracy on Training data : ', training_data_accuracy)

# accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

print('Accuracy on Test data : ', test_data_accuracy)
```

```
Accuracy on Training data :  0.9421487603305785
Accuracy on Test data :  0.819672131147541
```

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

model=RandomForestClassifier(n_estimators=50,max_depth=5)
model.fit(X_train,Y_train)
y_predict=model.predict(X_test)
y_pred_quant=model.predict_proba(X_test)[:,1]
y_pred=model.predict(X_test)

#evaluating model
print("Training Accuracy:",model.score(X_train,Y_train))
print("Testing Accuracy:",model.score(X_test,Y_test))

#classification report
cr=classification_report(Y_test,y_pred)
print(cr)
```

```
Training Accuracy: 0.9380165289256198
Testing Accuracy: 0.8032786885245902
              precision    recall  f1-score   support

           0       0.77      0.82      0.79        28
           1       0.84      0.79      0.81        33

    accuracy                           0.80        61
   macro avg       0.80      0.80      0.80        61
weighted avg       0.81      0.80      0.80        61
```



CONCLUSION

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Logistic Regression, Random Forest algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm 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.

REFERENCES

- Machine Learning Decision Tree Classification Algorithm - Javatpoint
- Ramalingam, V.V., Dandapath, A. and Raja, M.K., 2018. Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), pp.684-687.
- https://www.javatpoint.com/logistic-regression-in-machine-learnin