A

Major Project

On

HEART STROKE PREDICTION USING MACHINE LEARNING

(Submitted in partial fulfillment of the requirements for the award of Degree)

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND

ENGINEERING

BY

Sagar Yellaram(177R1A0555)

Sumanth Kothamasu(177R1A0526)

Surendharreddy Puchakayala(177R1A0544)

Under the Guidance of

B.P.Deepak Kumar

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR TECHNICAL CAMPUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi)
Recognized Under Section 2(f) & 12(B) of the UGC Act.1956,
Kandlakoya (V), Medchal Road, Hyderabad-501401.

2017-2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "**HEART STROKE PREDICTION**" being submitted by **Sagar Yellaram, Sumanth Kothamasu, Surendharreddy Puchakayala** bearing the **177R1A0555, 177R1A0526, 177R1A0544** roll number in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2020-21.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Mr.B.P.Deepak Kumar
Assistant Professor
INTERNAL GUIDE

Dr. A. Raji Reddy
DIRECTOR

Dr. K. Srujan Raju
HOD

EXTERNAL EXAMINER

Submitted for viva-voice Examination held on _____

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express our profound gratitude and deep regard to our guide

Mr.B.P.Deepak Kumar , Assistant Professor for her exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by her shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) Coordinators: **Ms. V.Sandhya, Mr. B.P.Deepak Kumar** and **Mr. J. Narasimha Rao** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to the Head of the Department **Dr. K. Srujan Raju** for providing excellent infrastructure and a nice atmosphere for completing this project successfully. We are obliged to our Director **Dr. A. Raji Reddy** for being cooperative throughout the course of this project. We would like to express our sincere gratitude to our Chairman Sri. **Ch. Gopal Reddy** for his encouragement throughout the course of this project The guidance and support received from all the members of **CMR TECHNICAL CAMPUS** who contributed and who are contributing to this project, was vital for the success of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement without which this assignment would not be possible. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project.

Sagar Yellaram(177R1A0555)

Sumanth Kothamasu(177R1A026)

Surendharreddy Puchakayala(177R1A0544)

ABSTRACT

In recent times, Heart Stroke prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart Stroke. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart stroke 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 model makes use of heart stroke dataset. The proposed work predicts the chances of Heart Stroke and classifies patient's risk level by implementing different data mining techniques such as KNN, Decision Tree and Random Forest.

LIST OF FIGURES

LIST OF SCREENSHOTS

TABLE OF CONTENTS

1. INTRODUCTION

1.INTRODUCTION

Machine learning is an application of artificial intelligence (AI) that pro-videos systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning in conjunction with big data can not only collect information but also find specific patterns. In recent times, Heart Stroke prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart Stroke. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart stroke 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 model makes use of heart stroke dataset. The proposed work predicts the chances of Heart Stroke and classifies patient's risk level by implementing different data mining techniques such as KNN, Decision Tree and Random Forest. Thus, this model presents a comparative study by analyzing the performance of different machine learning algorithms. The trial results verify that Random Forest algorithm has achieved the highest accuracy of 99.17% compared to other ML algorithms implemented.. The aim of this system is to provide a quick, immediate and easy way to predict accurately.

This system is exclusively for the health officials to easily predict heart stroke in well advance. The whole process of prediction is done by the trained model. This model helps the health professionals to utilize this system and predict early so that any person may not be affected severely. Over all this helps health department in a very efficient manner.

1.1 PROJECT PURPOSE

Here the purpose of this project is to train and test the data to Predict whether there is a chance for a person of getting a heart stroke or not. In this project we are going to predict the heart stroke data by using some machine learning algorithms. They are Decision Tree, KNN and Random Forest algorithms.

1.2 PROJECT SCOPE

The scope of this paper is to implement and investigate how different supervised binary classification methods impact default prediction. This is done by mining the data of the previous records of the people already available in kaggle website. So the idea of this project is to gather heart stroke data from multiple data sources and train machine learning algorithms on this data and to deploy this model we use flask framework which redirects it to a web browser that predicts whether there is chance of getting heart stroke or not.

1.3 PROJECT FEATURES

Data mining is the process of analyzing data from different perspectives and extracting useful knowledge from it. It is the core of knowledge discovery process. The various steps involved in extracting knowledge from the raw data. Different data mining techniques which it include classification, clustering, association the rule of mining, prediction and sequential patterns, neural networks, regression etc. Classification is the most commonly applied data mining technique which employs a set of classified examples to develop a model that can classify the population records at large Heart stroke can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart stroke can be reduced and the functioning of the heart improved. The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set Heart stroke prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart stroke. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive. This is done by mining the data of the previous records of the people. So the idea of this project is to gather heart stroke data from multiple data sources and train machine learning algorithms on this data and to deploy this model we use flask framework which redirects it to a web browser that predicts results.

2. SYSTEM ANALYSIS

2. SYSTEM ANALYSIS

SYSTEM ANALYSIS

System Analysis is the important phase in the system development process. The System is studied to the minute details and analyzed. The system analyst plays an important role of an interrogator and dwells deep into the working of the present system. In analysis, a detailed study of these operations performed by the system and their relationships within and outside the system is done. A key question considered here is, “what must be done to solve the problem?” The system is viewed as a whole and the inputs to the system are identified. Once analysis is completed the analyst has a firm understanding of what is to be done.

Requirement Specification is the part of the project which gives the details about the hardware and software requirements of the project. It also details the features of the programming language used. In order to implement this project, the choice of processor with maximum possible speed is made. There should be sufficient memory to store data and software tools for efficient processing.

2.1 PROBLEM DEFINITION

A detailed study of the process of heart stroke prediction using various machine learning algorithms. The data can be collected from various sources. First we have to take data set from health sectors, hospitals which has various details about the heart patients such as such as patient personal details, age, gender, BMI, average glucose level, smoking status, heart disease status, hyper tension , work type, etc. Patients first apply for a doctor appointment after that the doctor validates the patient health status. However doing this manually takes a lot of time because the patient needs to undergo various tests and wait until the reports arrive. Hence it wants to automate the heart stroke prediction(real time) based on patient information.

So the final thing is to identify the persons that are possible for being attacked by a heart stroke. How will the health sector, patient get benefited if we develop this model is the immediate question that arises. The solution is that this model is developed to predict well in advance to avoid risk and help patient taking suitable measures by consulting a doctor.

Hence the more accurate we are in predicting the person's chances of getting heart stroke by training the data with machine learning algorithms and deploying this model we use the flask framework which redirects it to a web browser that predicts whether he/she suffers from heart stroke or not.

2.2 EXISTING SYSTEM

Many different predicting systems are available but they make use of dataset and predict the result by taking the input parameters Lot of work has been carried out to predict heart stroke using UCI Machine Learning dataset. Different levels of accuracy have been attained using various data mining techniques . But good accuracy is not achieved.

2.2.1 LIMITATIONS OF EXISTING SYSTEM

- Accuracy is not better
- Lack in providing good results and user friendly interface.
- Incorrect Classification results.

2.3 PROPOSED SYSTEM

The primary goal of this project is to extract patterns from a common heart stroke data set, and then build a model based on these extracted patterns, in order to predict the likely of getting heart stroke by using classification data mining algorithms. Firstly data is collected and then data preprocessing is done and later data is split into train data and test data. Using train data the different type of Machine Learning models are trained and accuracy is predicted as a result and Plotting a graph. The objective of this study is to effectively predict if the patient suffers from heart stroke. The data is fed into model which predicts the probability of having heart stroke.

2.3.1 ADVANTAGES OF THE PROPOSED SYSTEM

- User friendly interface
- Efficient and Accurate
- Very helpful in predicting based on given inputs.
- High Performance

2.4 SYSTEM REQUIREMENTS SPECIFICATION

2.4.1 FUNCTIONAL REQUIREMENTS

- Input data to train the model.
- Data processing to clean the raw input data.
- Accurate algorithm for training model.
- Post-processing to manage the output.

2.4.2 NON-FUNCTIONAL REQUIREMENTS

- **Accuracy:** Accuracy is the major part as we require accurate results.
- **Reliability:** Reliability is needed so that model is not disturbed.
- **Adaptability**: The model should adapt automatically according to user input.
- **Security:** Security has to be provided to the data model in case of sensitive data.
- **Feasibility Study:** The model to be trained on data has to be feasible.

2.5 HARDWARE & SOFTWARE REQUIREMENTS

2.5.1 SOFTWARE REQUIREMENTS

Software Requirements specifies the logical characteristics of each interface and software components of the system. The following are some software requirements,

- Operating system : Windows 8, 10
- Languages : Python
- Back end : Machine Learning
- IDE : Jupyter

2.5.2 HARDWARE REQUIREMENTS

Hardware interfaces specify the logical characteristics of each interface between the software product and the hardware components of the system. The following are some hardware requirements.

- Processor : intel i3 or above.
- Hard disk : 16GB and Above.
- RAM : 4GB and Above.
- Monitor : 5 inches or above.

3. ARCHITECTURE

3.ARCHITECTURE

3.1 PROJECT ARCHITECTURE

This project architecture shows the procedure for loan approval prediction using machine learning Algorithms, starting from input to final prediction.

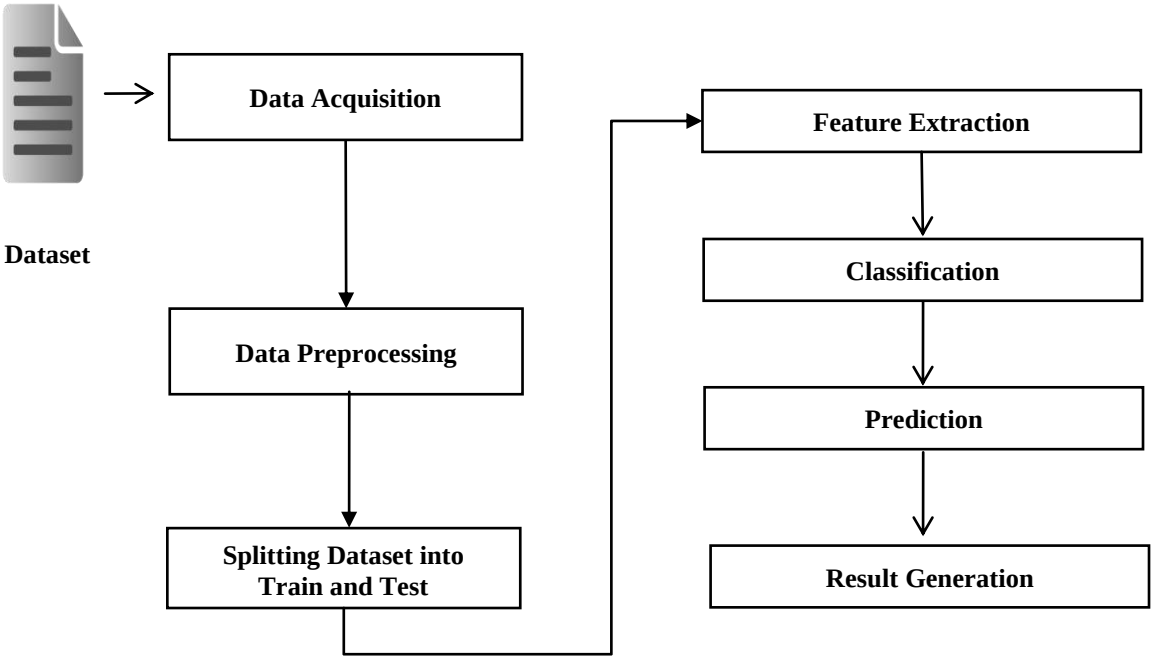


Figure 3.1: Project Architecture

3.1.1 Description

This model is prepared by using different modules and each module has strategy, some certain functionality and character in building up the model .

The modules present in this model are:

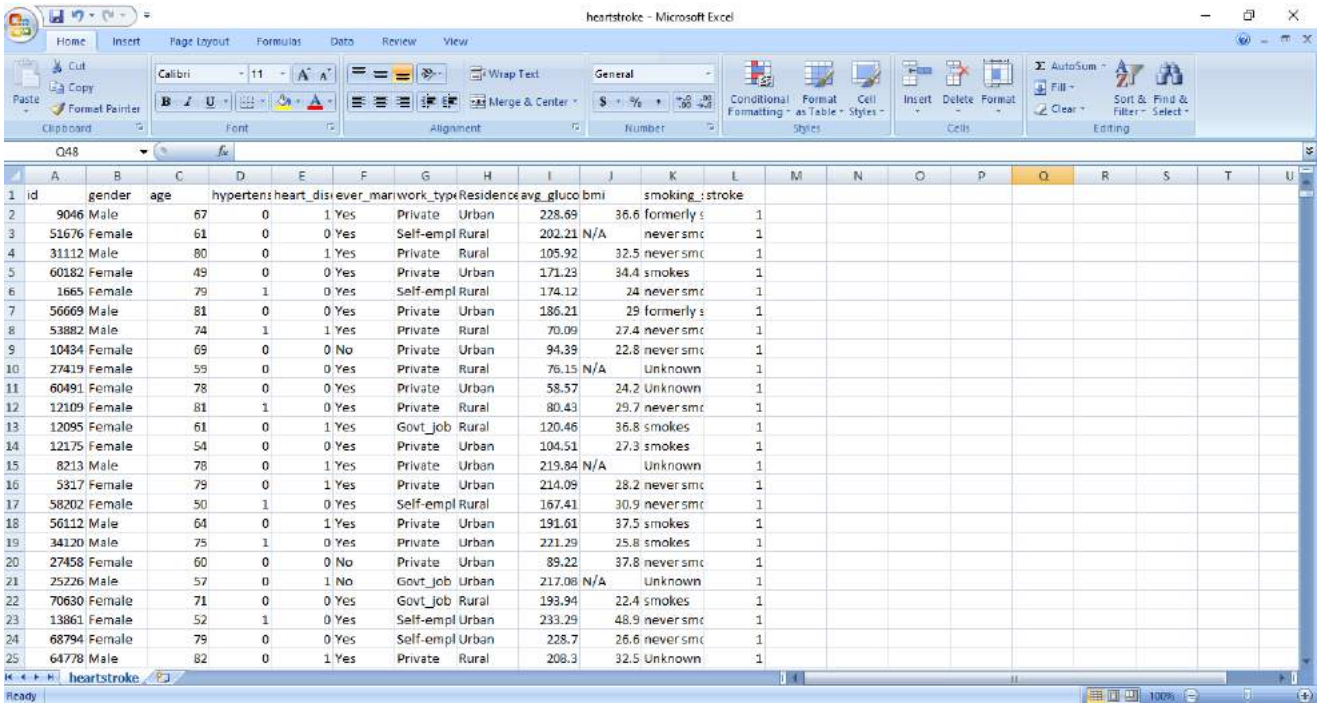
- Data Acquisition
- Data Preprocessing
- Data Analysis and Splitting data
- Feature Extraction
- Training using classifiers
- Testing model with classifiers
- Result Generation

3.1.2 Data Acquisition

As machine learning is based on available data for the system to make a decision hence the first step defined in the architecture is data acquisition. This involves data collection, preparing and segregating the case scenarios based on certain features involved with the decision making cycle and forwarding the data to the processing unit for carrying out further categorization.

We collected the data that is available in the kaggle website (i.e. heart stroke dataset).It is a collection of nearly 5000+ records of patients. The data selection is the process of selecting the data for detecting the attacks. In this project, the Heart stroke dataset is used for detecting heart stroke. The dataset which contains the information like id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke. The above mentioned are the attributes that are given as input to the model that is built using classifiers.

The attributes need to be preprocessed further in the next stage i.e. in the data preprocessing phase.



id	gender	age	hypertens	heart_dis	ever_mar	work_typ	Residence	avg_gluco	bmi	smoking_s	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly s	1
51676	Female	61	0	0	Yes	Self-empl	Rural	202.21	N/A	never smc	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smc	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-empl	Rural	174.12	24	never smc	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly s	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smc	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smc	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smc	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smc	1
58202	Female	50	1	0	Yes	Self-empl	Rural	167.41	30.9	never smc	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smc	1
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-empl	Urban	233.29	48.9	never smc	1
68794	Female	79	0	0	Yes	Self-empl	Urban	228.7	26.6	never smc	1
64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1

Figure 3.1.2 : Heart Stroke Dataset

3.1.3 Data Preprocessing

Data pre-processing is the process of removing the unwanted data from the dataset. The received data in the data acquisition layer is then sent forward to the data preprocessing layer where it is subjected to advanced integration and processing and involves data cleaning and encoding.

- ✓ Missing data removal
- ✓ Encoding Categorical data

Missing data removal: In this process, the null values such as missing values are removed using imputer library.

Encoding Categorical data: That categorical data is defined as variables with a finite set of label values. That most machine learning algorithms require numerical input and output variables. That an integer and standard encoding is used to convert categorical data to integer data.

3.1.4 Data Analysis and Splitting Data

The data set that uses EDA undergoes the process of normalization, null value treatment, choosing essential columns using filtering, identifying the target variables and visualizing the data in the graphical format. Python is used for easy and efficient processing of data. We used the pandas library available in Python to process and extract information from the given data set. Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes. One portion of the data is used to develop a predictive model. And the other to evaluate the model's performance. Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.

3.1.5 Feature Extraction

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. Standardization is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm

3.1.6 Training using Classifiers

In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called empirical risk minimization. Loss is the penalty for a bad prediction. That is, loss is a number indicating how bad the model’s prediction was on a single example. If the model’s prediction is perfect, the loss is zero; otherwise, the loss is greater. The goal of training a model is to find a set of weights and biases that have low loss, on average, across all examples. Here we have used three algorithms -

- Decision Tree
- KNN
- Random Forest

Training the model with different algorithms

1.Decision Tree

```
In [49]: from sklearn.tree import DecisionTreeClassifier
         dt=DecisionTreeClassifier()

In [50]: dt.fit(X_train_std,Y_train)
Out[50]: DecisionTreeClassifier()

In [51]: dt.feature_importances_
Out[51]: array([0.01712356, 0.418615  , 0.02644392, 0.01354881, 0.01970826,
                0.04021249, 0.00775032, 0.23077888, 0.17877853, 0.04704023])

In [52]: X_train.columns
Out[52]: Index(['gender', 'age', 'hypertension', 'heart_disease', 'ever_married',
                'work_type', 'Residence_type', 'avg_glucose_level', 'bmi',
                'smoking_status'],
               dtype='object')
```

Figure 3.1.6.1: Training the model with Decision Tree

2.KNN

```
In [60]: from sklearn.neighbors import KNeighborsClassifier
         knn=KNeighborsClassifier()

In [61]: knn.fit(X_train_std,Y_train)
Out[61]: KNeighborsClassifier()
```

Figure 3.1.6.2: Training the model with KNN

3.Random Forest

```
In [67]: from sklearn.ensemble import RandomForestClassifier
         rf=RandomForestClassifier()

In [68]: rf.fit(X_train_std,Y_train)

Out[68]: RandomForestClassifier()
```

Figure 3.1.6.3: Training the model with Random Forest

3.1.7 TESTING

1.Decision Tree

```
In [53]: Y_pred_dt=dt.predict(X_test_std)

In [54]: Y_pred_dt
Out[54]: array([1, 1, 0, ..., 0, 0, 0], dtype=int64)

In [55]: from sklearn.metrics import accuracy_score

In [56]: ac_dt=accuracy_score(Y_test,Y_pred_dt)

In [57]: ac_dt
Out[57]: 0.9629883481836875

In [58]: from sklearn import metrics
         cm=metrics.confusion_matrix(Y_test,Y_pred_dt)
         print(cm)
         [[905  54]
          [  0 500]]
```

Figure 3.1.7.1 : Testing the model using Decision Tree algorithm

2.KNN

```
In [62]: Y_pred_knn=knn.predict(X_test_std)

In [63]: ac_knn=accuracy_score(Y_test,Y_pred_knn)

In [64]: ac_knn
Out[64]: 0.9047292666209733

In [65]: from sklearn import metrics
         cm=metrics.confusion_matrix(Y_test,Y_pred_knn)
         print(cm)
         [[820 139]
          [  0 500]]
```

Figure 3.1.7.2: Testing the model using KNN algorithm

3.Random Forest

```
In [69]: Y_pred_rf=rf.predict(X_test_std)

In [70]: ac_rf=accuracy_score(Y_test,Y_pred_rf)

In [71]: ac_rf

Out[71]: 0.9910897875257025

In [74]: from sklearn import metrics
         cm=metrics.confusion_matrix(Y_test,Y_pred_rf)
         print(cm)

         [[946  13]
          [  0 500]]
```

Figure 3.1.7.3: Testing the model using Random Forest algorithm

3.1.8 RESULT AS A GRAPH

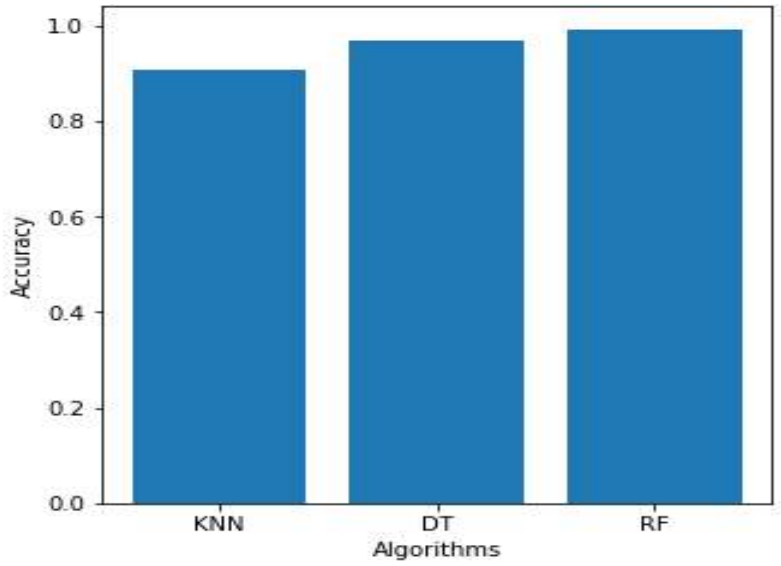


Figure 3.1.8: Predicting the Result

3.2 USE CASE DIAGRAM

Use case diagrams represent the functionality of the system from a user point of view. A Use case describes a function provided by the system that yields a visible result for an actor. An actor describes any entity that interacts with the system.

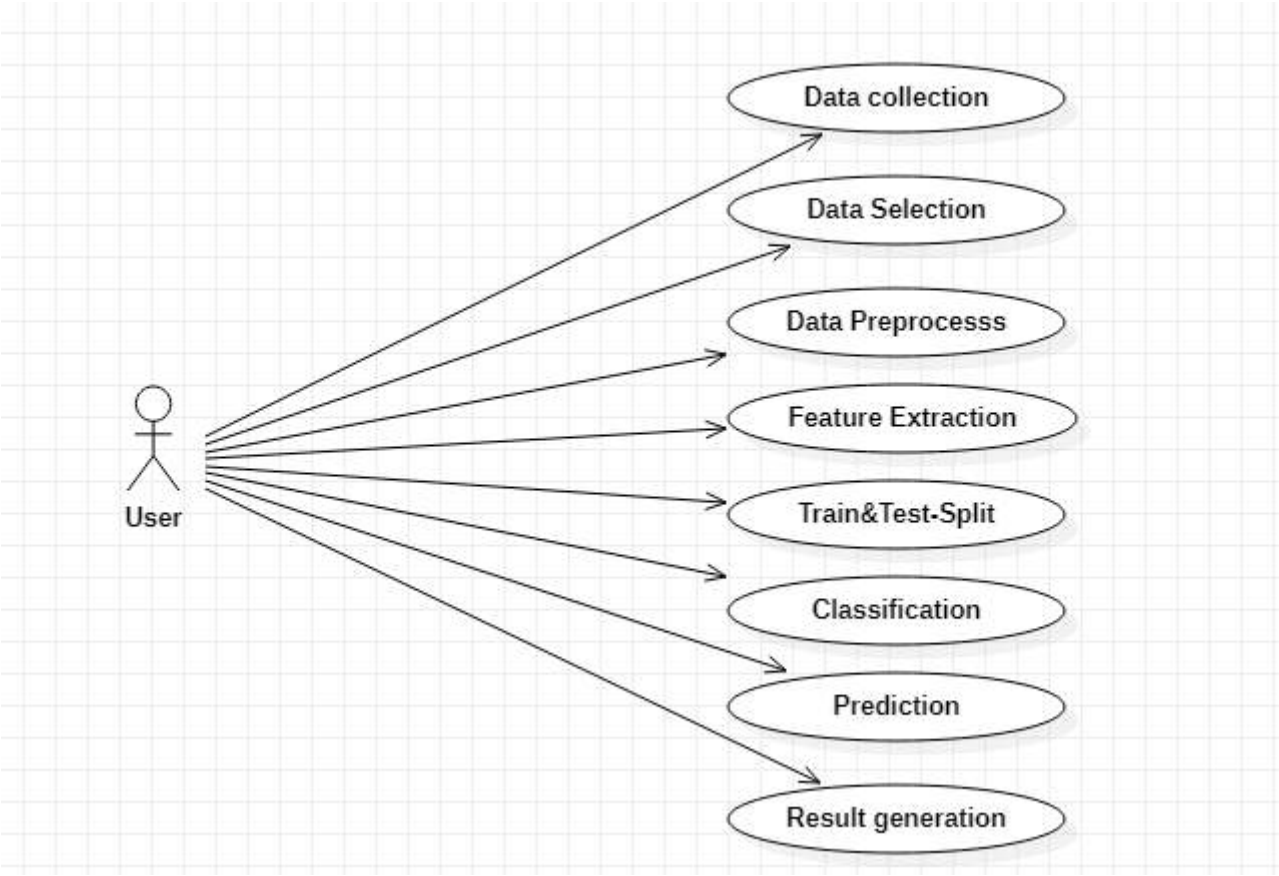


Figure 3.2: Use Case Diagram for user for Heart Stroke Prediction System

3.3 SEQUENCE DIAGRAM

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

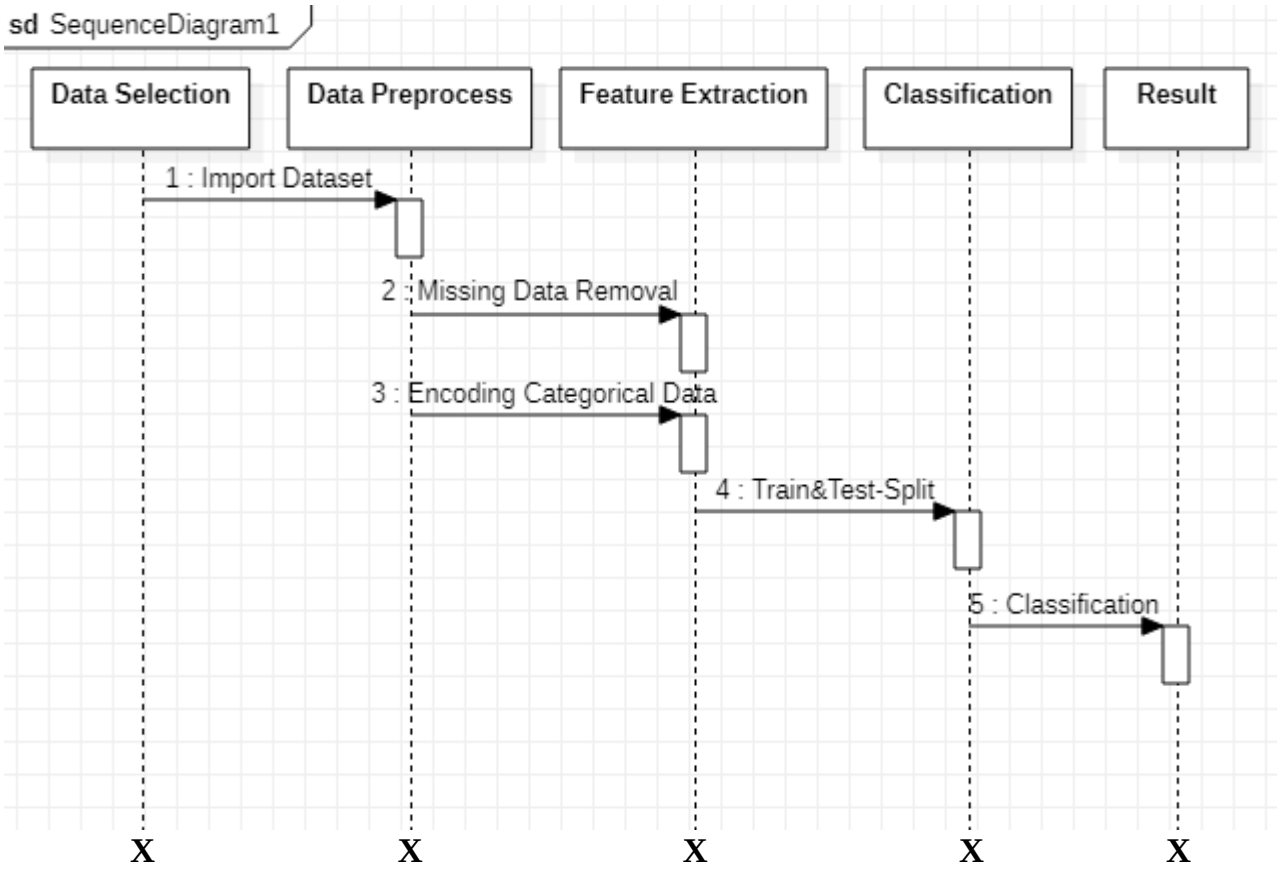


Figure 3.3: Sequence Diagram for Heart Stroke Prediction

3.4 CLASS DIAGRAM

Class Diagram is a collection of classes and objects. The class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing and documenting different aspects of a system but also for constructing executable code of the software application. The class diagram shows a collection of classes, interfaces, associations, collaborations and constraints.

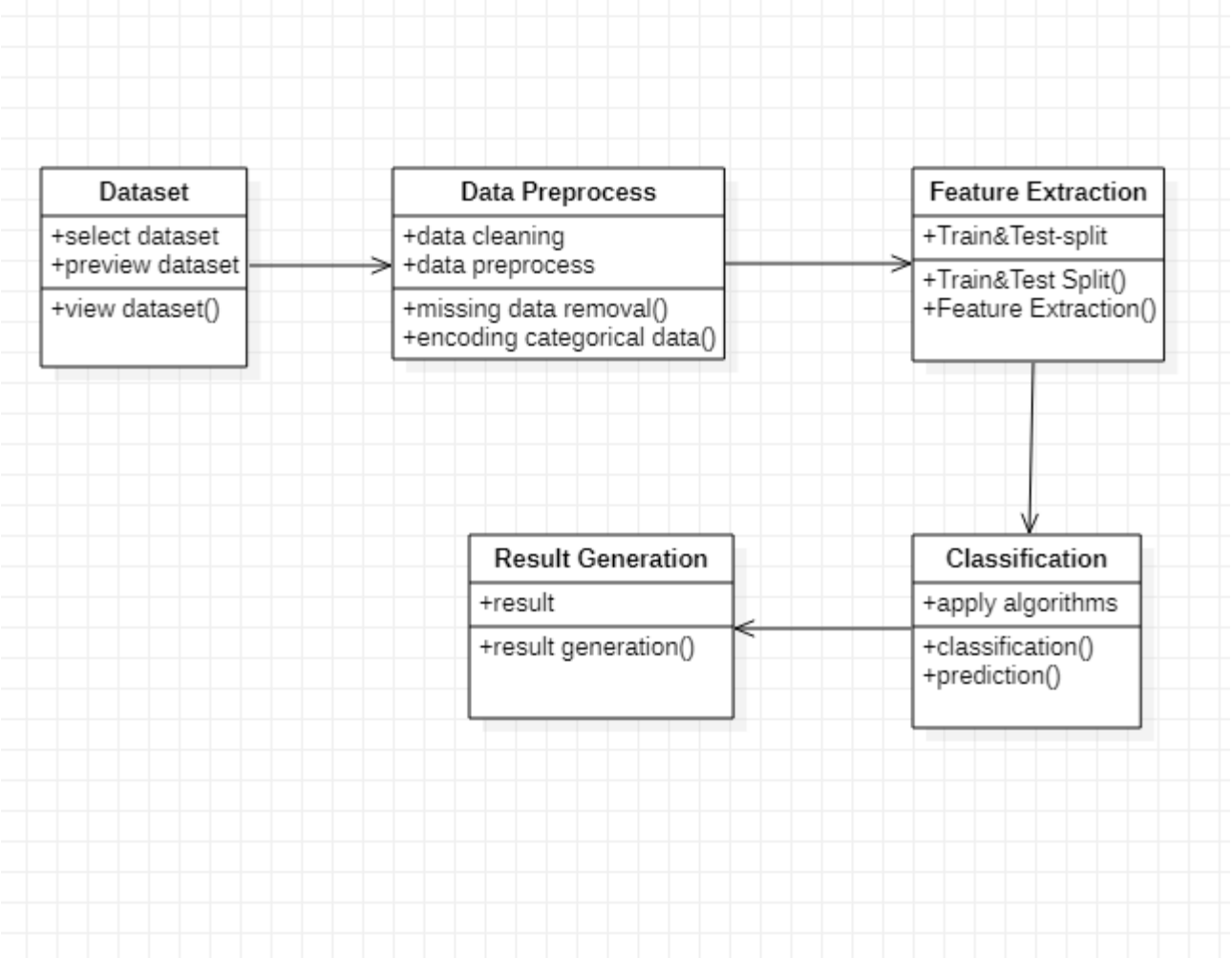


Figure 3.4: Class Diagram for Heart Stroke Prediction

3.5 ACTIVITY DIAGRAM

It describes the flow of activity states. A collaboration diagram is a type of visual presentation that shows how various software objects interact with each other within an overall IT architecture and 30 how users can benefit from this collaboration. A collaboration diagram often comes in the form of a visual chart that resembles a flow chart.

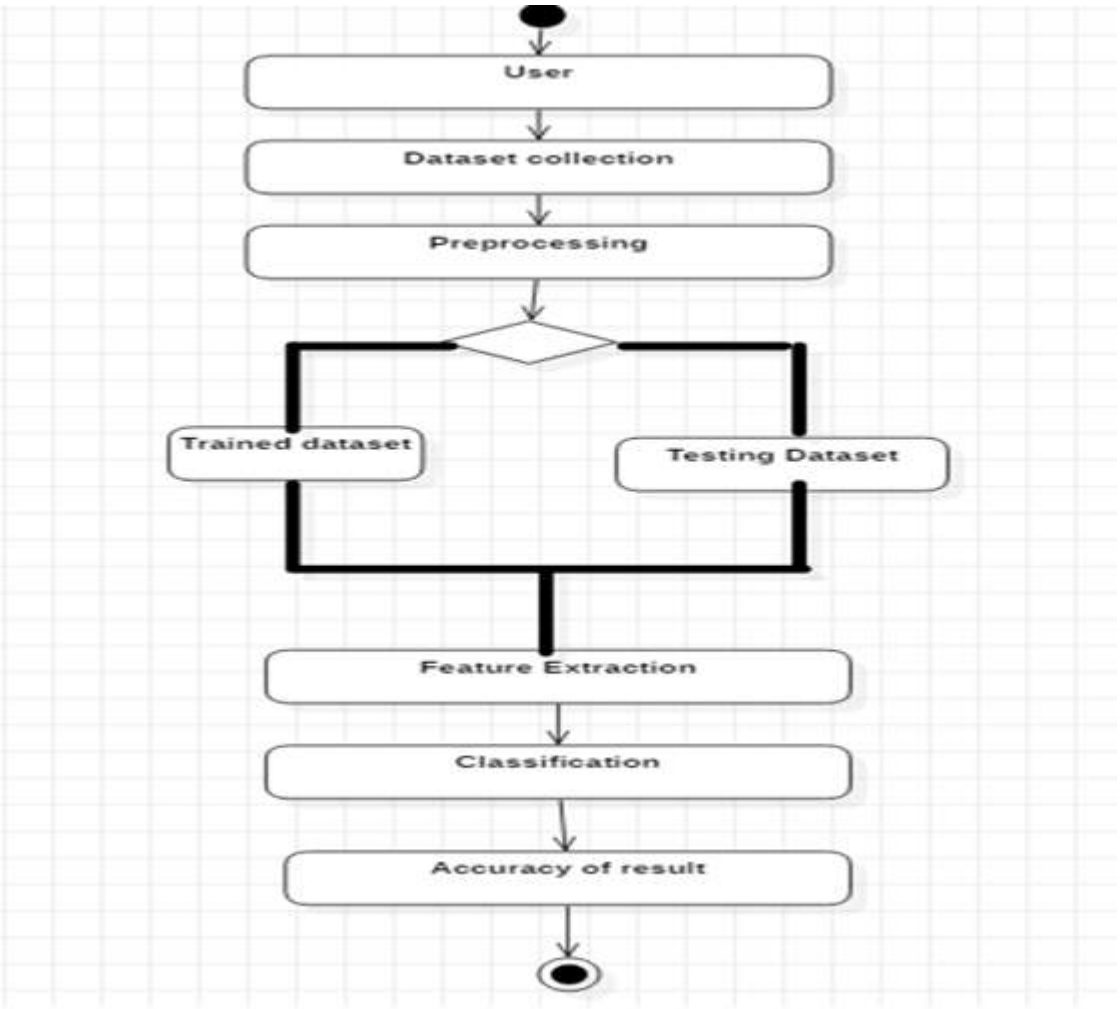


Figure 3.5: Activity Diagram for Heart Stroke Prediction

4. IMPLEMENTATION

SAMPLE CODE

```
from flask import Flask, render_template, request
import joblib
import os
import  numpy as np
import pickle
app= Flask(__name__)

@app.route("/")
def index():
   return render_template("home.html")
@app.route("/result",methods=['POST','GET'])
def result():
   gender=int(request.form['gender'])
   age=int(request.form['age'])
   hypertension=int(request.form['hypertension'])
   heart_disease = int(request.form['heart_disease'])
   ever_married = int(request.form['ever_married'])
   work_type = int(request.form['work_type'])
   Residence_type = int(request.form['Residence_type'])
   avg_glucose_level = float(request.form['avg_glucose_level'])
   bmi = float(request.form['bmi'])
   smoking_status = int(request.form['smoking_status'])

   x=np.array([gender,age,hypertension,heart_disease,ever_married,work_type,Residence_type,
           avg_glucose_level,bmi,smoking_status]).reshape(1,-1)

   scaler_path=os.path.join('D:/majorproject','models/scaler.pkl')
   scaler=None
   with open(scaler_path,'rb') as scaler_file:
      scaler=pickle.load(scaler_file)

   x=scaler.transform(x)

   model_path=os.path.join('D:/majorproject','models/rf.sav')
   rf=joblib.load(model_path)

   Y_pred=rf.predict(x)

   # for No Stroke Risk
   if Y_pred==0:
      return render_template('noheartstroke.html')
   else:
      return render_template('heartstroke.html')

if __name__=="__main__":
   app.run(debug=True,port=7384)
```

5. SCREENSHOTS

5.1 HOME PAGE

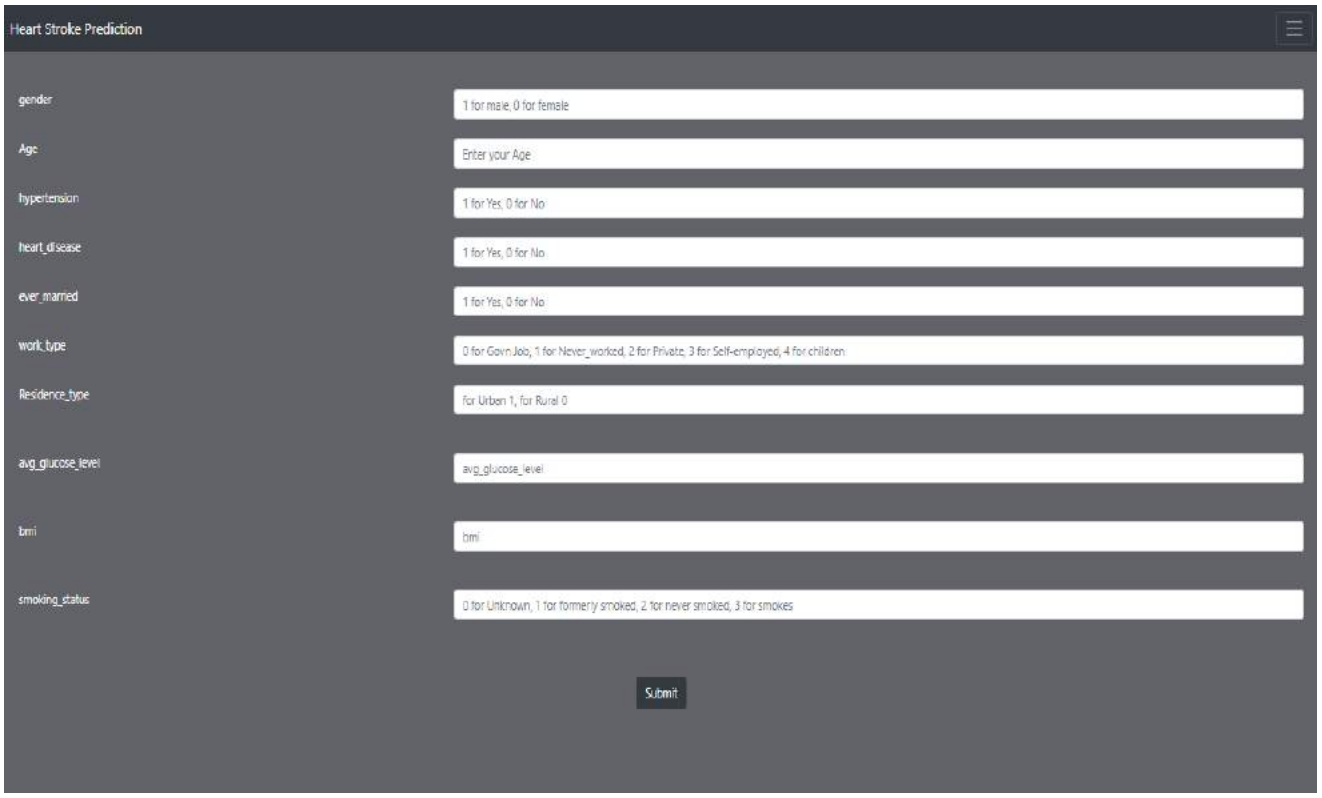


Figure 5.1 Home Page of Heart Stroke Prediction

5.2 PREDICTION OF HEART STROKE PAGE

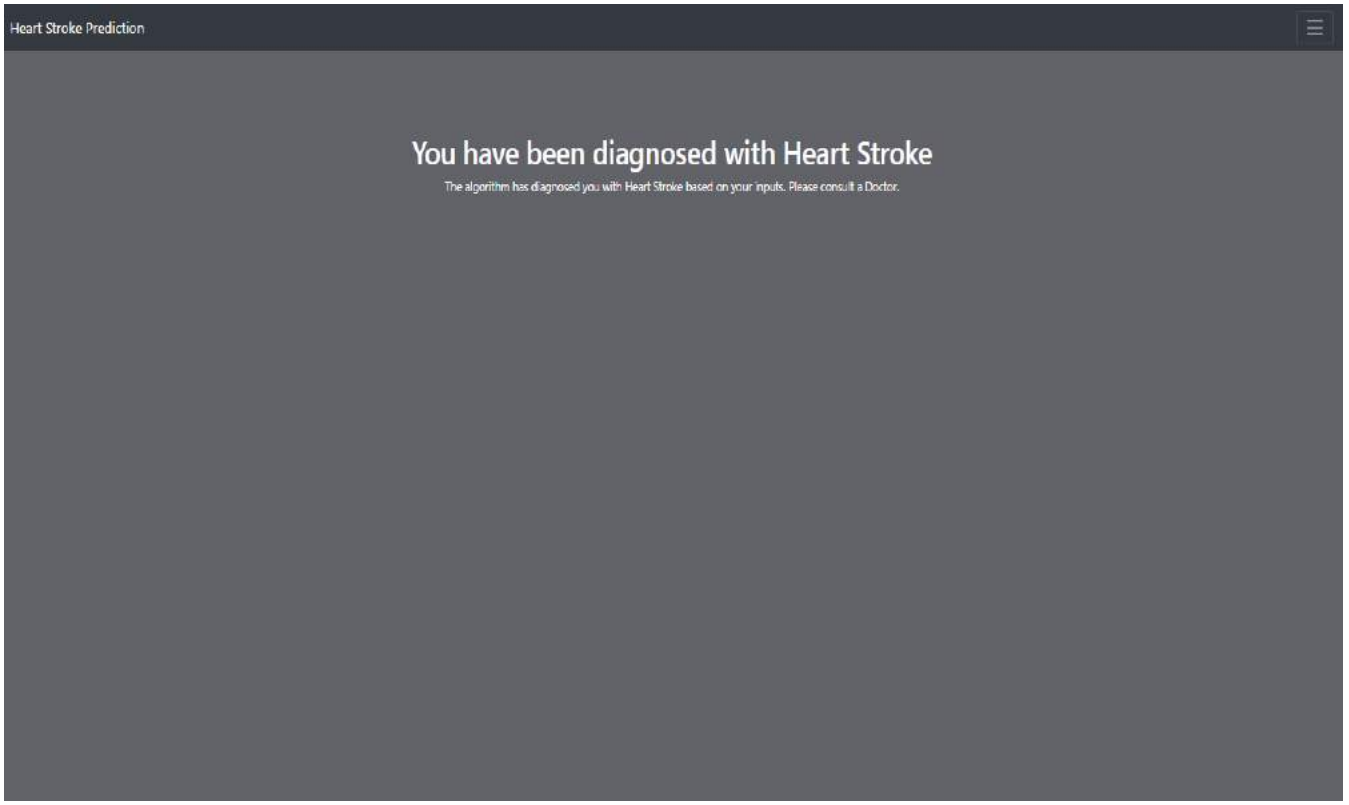


Figure 5.2 Prediction of Heart Stroke Page

5.3 PREDICTION OF NO HEART STROKE PAGE

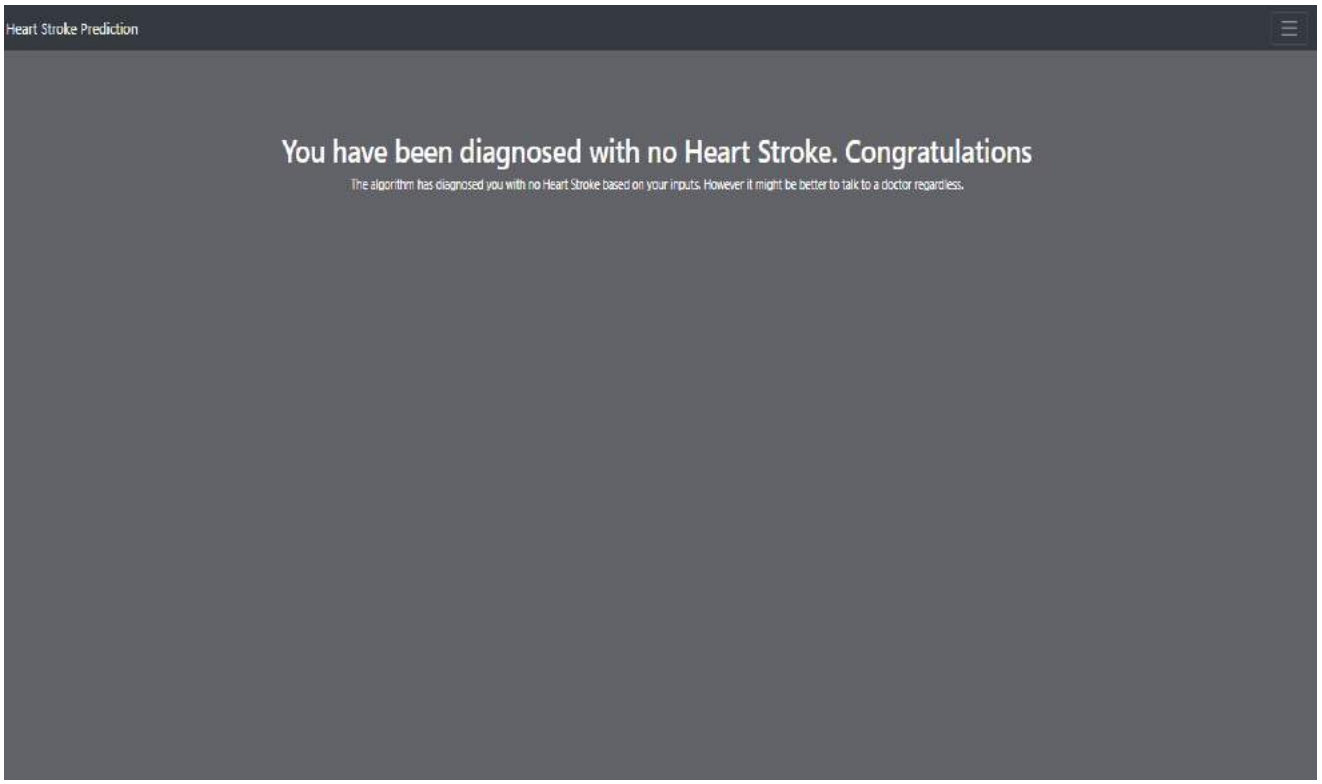


Figure 5.3 Prediction of No Heart Stroke Page

6. TESTING

6. TESTING

6.1 INTRODUCTION TO TESTING

Testing is the process of finding differences between the expected behavior specified by system models and the observed behavior Testing is the process of executing a program with the intent of finding any errors. Testing is vital to the success of the system. Without proper testing hide errors will surface after sometime of use and perhaps irreversible damage has been done to valuable data. A series of tests like responsiveness, its value, stress and security are performed before the system is ready for user acceptance testing. System testing follows the logical conclusion that all parts of the system are tested and found to be working properly under all kinds of situations, and then the system is achieving its goal of processing the data perfectly according to user rules and requirements.

6.2 TESTING ACTIVITIES

Different levels of testing are used in the testing process, each level of testing aims to test different aspects of the system. The basic levels are:

Unit testing

Integration testing

System testing

Acceptance testing

6.2.1 UNIT TESTING

Unit testing focuses on the building blocks of the software system, that is, objects and sub system. There are three motivations behind focusing on components. First, unit testing reduces the complexity of the overall tests activities, allowing us to focus on smaller units of the system. Second, unit testing makes it easier to pinpoint and correct faults given that few components are involved in the test. Third, Unit testing allows parallelism in the testing activities in which each component can be tested independently and then we combine them and perform integration testing.

6.2.2 INTEGRATION TESTING

In the integration testing, many test modules are combined into sub systems, which are then tested. The goal here is to see if the modules can be integrated properly, the emphasis being on testing module interaction. After structural testing and functional testing get error free modules . These modules are to be integrated to get the required results of the system. After checking a module, another module is tested and is integrated with the previous module. After the integration, the test cases are generated and the results are tested.

6.2.3 SYSTEM TESTING

In system testing the entire software is tested. The reference document for the process is the requirement document and the goal is to see whether the software meets its requirements. The system was tested for various test cases with various inputs.

6.2.4 ACCEPTANCE TESTING

Acceptance testing is sometimes performed with realistic data of the client to demonstrate that the software is working satisfactory. Testing here focuses on the external behaviour .

6.3 TYPES OF TESTING

- Black box or functional testing
- White box testing or structural testing

6.3.1 BLACK BOX TESTING

The method is used when knowledge of the specified function that a product has been designed to perform is known. The concept of black box is used to represent a system whose inside workings are not available to inspection. In a black box the test item is a "Black" , since its logic is unknown , all that is known is what goes in and what comes out or the input and output.

6.3.2 WHITE BOX TESTING

White box testing is concerned with testing the implementation of the program. The intent of structural is not to exercise all the inputs or outputs but to exercise the different programming and data structure used in the program. Thus structural testing aims to achieve test cases that will force the desired coverage of different structures. Two types of path testing are statement testing coverage and branch testing coverage.

Test Plan:

Testing process starts with a test plan. The plan identifies all the testing related activities that must be performed and specifies the schedules, allocates the resources, and specified guidelines for testing. During the testing of the unit the specified test cases are executed and the actual result compared with expected output. The final output of the testing phase is the test report and the error report.

Test Data:

Here all test cases that are used for the system testing are specified. The goal is to test the different functional requirements specified in Software Requirements Specifications (SRS) document

Test Report:

The module is working properly provided the user has to enter information. All data entry forms have been tested with specified test cases and all data entry forms are working properly.

6.4 TEST CASES

We have used three algorithms and trained it so that we can choose the best algorithm for Heart Stroke Prediction.

6.4.1.TESTING AT FIRST RUN

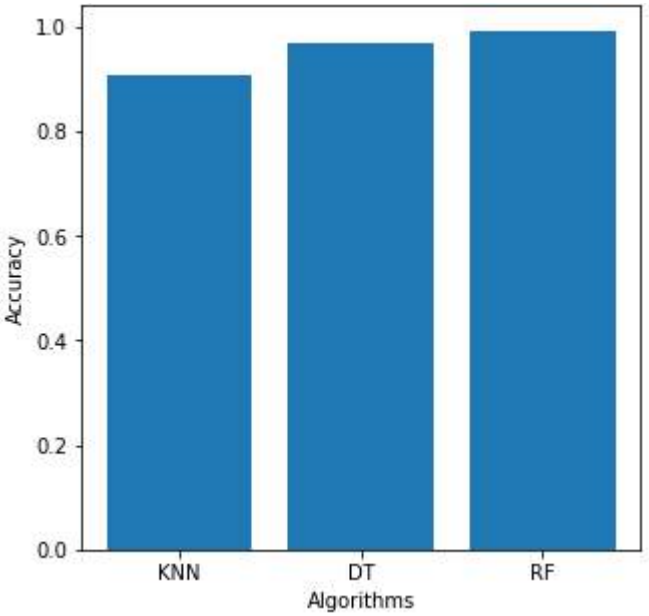


Figure 6.4.1: Testing at first run

6.4.2 TESTING AT SECOND RUN

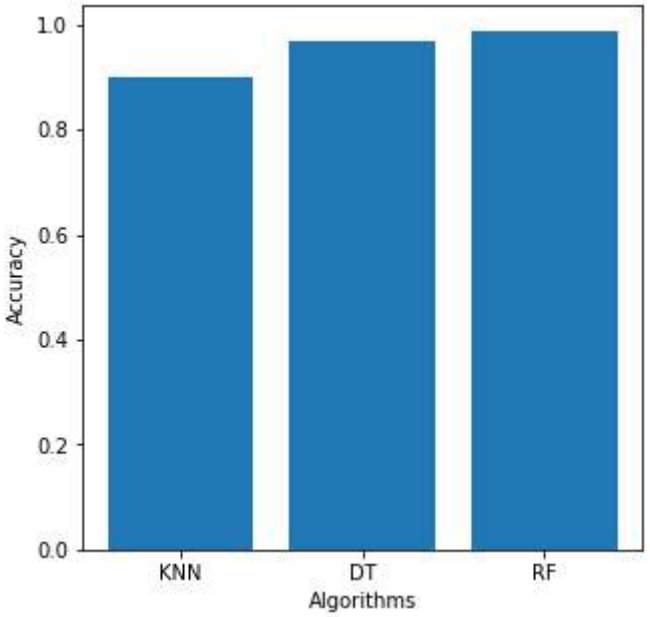


Figure 6.4.2: Testing at second run

6.4.3 TESTING AT THIRD RUN

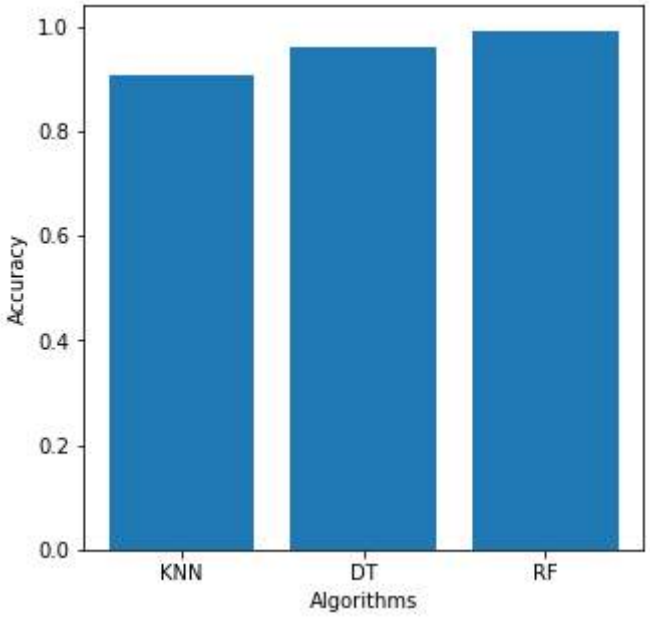


Figure 6.4.3: Testing at third run

6.4.4 TESTING AT FOURTH RUN

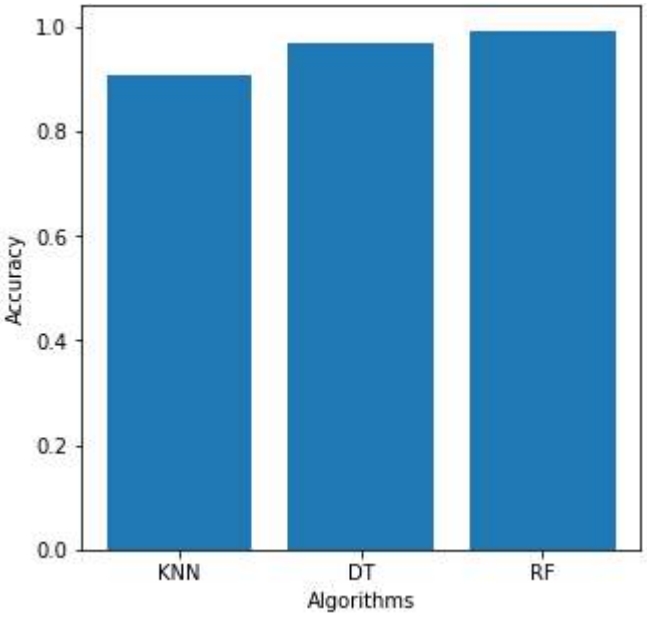


Figure 6.4.4: Testing at fourth run

6.4.5 ACCURACIES OF EACH ALGORITHM IN DIFFERENT RUNS

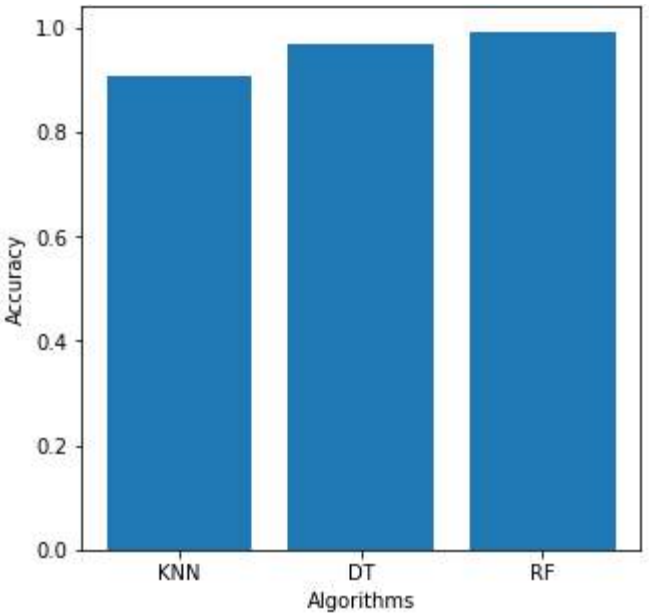


Figure 6.4.5: Accuracies of Each algorithm in different runs

By comparing the accuracy of all the three algorithms by running and testing them various times, we can say that Random Forest has the better accuracy than other algorithms in all test cases.

7. CONCLUSION

CONCLUSION

7.1 PROJECT CONCLUSION

The machine learning field is continuously evolving. And along with evolution comes a rise in demand and importance. There is one crucial reason why data scientists need machine learning, and that is: 'High-value predictions that can guide better decisions and smart actions in real-time without human intervention.'

The proposed model is GUI based, scalable, user-friendly, reliable and expandable system. The proposed model can also help in reducing treatment costs by providing initial diagnosis in time. This model can also serve as a purpose for training medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio patients. With the increasing number of deaths due to heart stroke, it has become mandatory to develop a system to predict heart stroke effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart stroke. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 99.17% for prediction of heart stroke. 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 stroke effectively and efficiently.

7.2 FUTURE ENHANCEMENTS

In future we can use this system for analysis of different data sets. The dataset analysed in this study was from a single network. This study can be conducted with a larger and smaller network area. This would allow determining if the readmission factors differ based on patient geographical location or if similar traits are observed nationwide. In addition, this would strengthen both urban and rural models while assessing the importance of age categorization. In near future Healthcare systems are moving towards Analytics and Data network to find the Prediction through machine learning or deep learning. Identifying and selection of significant attributes for better diagnosis of heart stroke are very challenging tasks for future research. There are many possible improvements that could be explored to improve scalability and accuracy of this prediction system.

8. BIBLIOGRAPHY

8.BIBLIOGRAPHY

8.1 GITHUB REPOSITORY LINK

htttps://github.com/sumanth2522/Heart-Stroke-Prediction-using-ML

8.2 REFERENCES

1. L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi, and N. Toghianfar, “Prediction and control of heart stroke by data mining,” International Journal of Preventive Medicine, vol. 4, no. Suppl 2, pp. S245–249, May 2013.
2. S.-F. Sung, C.-Y. Hsieh, Y.-H. Kao Yang, H.-J. Lin, C.-H. Chen, Y.- W. Chen, and Y.-H. Hu, “Developing a heart stroke severity index based on administrative data was feasible using data mining techniques,” Journal of Clinical Epidemiology, vol. 68, no. 11, pp. 1292–1300, Nov. 2015.
3. P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, “Classification of heart stroke disease using machine learning algorithms,” Neural Computing and Applications, vol. 32, no. 3, pp. 817–828, Feb. 2020.
4. H.M. Nguyen, E.W. Cooper, K. Kamei ,A comparative study on sampling Techniques for handling class imbalance in streaming data, SCIS-ISIS 2012.

8.3 WEBSITES

- https://www.flask.com
- https://www.machinelearningexpert.com
- https://www.kaggle.com
- https://www.datascience.com
- https://www.anaconda.com
- https://www.pythonanywhere.com
- https://www.github.com

IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Heart Stroke Prediction using Machine Learning

B.P. Deepak Kumar[*1], Sagar Yellaram[*2], Sumanth kothamasu[*3], SurendharReddy Puchakayala[*4]

[*1]Assistant Professor,Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

[*2]JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

[*3]JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

[*4]JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

ABSTRACT

In recent times, Heart Stroke prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart Stroke. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart stroke 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 stroke dataset. The proposed work predicts the chances of Heart Stroke and classifies patient's risk level by implementing different data mining techniques such as KNN, Decision Tree and Random Forest. Thus, this paper presents a comparative study by analyzing the performance of different machine learning algorithms. The trial results verify that Random Forest algorithm has achieved the highest accuracy of 99.17% compared to other ML algorithms implemented.

Keywords: **KNN, Decision Tree, Random Forest, Heart Stroke Prediction.**

I. INTRODUCTION

The work proposed in this paper focus mainly on various data mining practices that are employed in heart stroke Prediction. Human heart is the principal part of the human body. Any irregularity to heart can cause distress in other parts of body. In today's contemporary world, heart stroke is one of the primary reasons for occurrence of most deaths. Heart stroke may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hyper-tension . According to the World Health Organization more than 10 million die due to Heart stroke every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart stroke. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis . The proposed work makes an attempt to detect these heart stroke at early stage to avoid disastrous consequences. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart stroke 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 analyze the given data. After analysis of data ML techniques help in heart stroke prediction and early diagnosis. This paper presents performance analysis of various ML techniques such as KNN, Decision Tree, and Random Forest for predicting heart stroke at an early stage

II. LITERATURE SURVEY

Many researchers have already used machine learning based approached to predict heart strokes. Govindarajan et al. [11] conducted a study to categorize heart stroke disorder using a text mining combination and a machine learning classifier and collected data for 507 patients. For their analysis, they used various machine learning approaches for training purposes using ANN, and the SGD algorithm gave them the best value, which was 95%. Amini et al. [4], [12] conducted research to predict stroke incidence, collected 807 healthy and unhealthy subjects in their study categorized 50 risk factors for stroke, diabetes, cardiovascular disease, smoking, hyperlipidemia, and alcohol use. They used two techniques that had the best accuracy from c4.5 decision tree algorithm, and it was 95%, and for K-nearest neighbor, the accuracy was 94%. Cheng et al. [13] published a report on the estimation of the heart stroke prognosis. In their analysis, 82 stroke patient data were used, two ANN models were used to find precision, and 79% and 95% were used. Cheon et al. [14]–[16] performed a study to predict stroke patient mortality. In their study, they used 15099 patients to identify heart stroke occurrence. They used a deep neural network approach to detect heart strokes. The authors used PCA to extract medical record history and predict heart stroke. They have got an area under the curve (AUC) value of 83%. Singh et al.

[17] performed a study on heart stroke prediction applied to artificial intelligence. In their research, they used a different method for predicting stroke on the cardiovascular health study (CHS) dataset. And they took the decision tree algorithm to feature extract to principal component analysis. They used a neural network classification algorithm to construct the model they got 97% accuracy. Chin et al. [18] performed a study to detect an automated early heart stroke. In their study, the main purpose was to develop a system using CNN to automated primary heart stroke. They collected 256 images to train and test the CNN model. In their system image prepossessing remove the impossible area that can't occur of heart stroke, they used the data prolongation method to raise the collected image. Their CNN method has given 90% accuracy.

The main idea behind the proposed system after reviewing the above papers was to create a heart stroke prediction system based on the inputs . We analysed the classification algorithms namely KNN, Decision Tree and Random Forest based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction.

III.PROPOSED SYSTEM

The proposed work predicts heart stroke by exploring the above mentioned four classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart stroke. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart stroke. Fig. 1 shows the entire process involved .

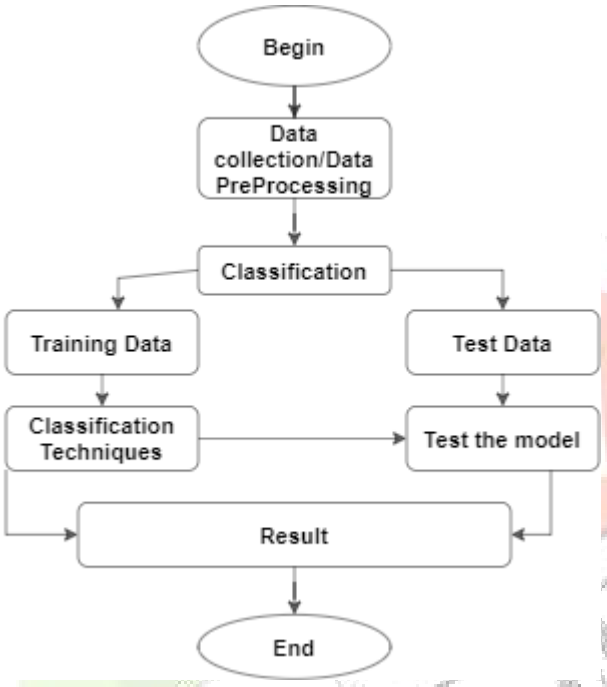


Fig 1: Generic Model Predicting Heart Stroke

METHODOLOGY:

This section is divided into two parts, these are: Data description, machine learning classifiers . These two processes are described below:

A) Data Description:

Here we used the heart stroke dataset that is available in the kaggle website for our analysis. This dataset consists of total 12 attributes. The complete description of the attributes used in the proposed work is given below:

id : This attribute means person's id. It's numerical data.
Age : This attribute means a person's age. It's numerical data.
Gender : This attribute means a person's gender. It's categorical data.
Hypertension : This attribute means that this person is hypertensive or not. It's numerical data.
work type : This attribute represents the person work scenario. It's categorical data.
residence type : This attribute represents the person living scenario. It's categorical data.
heart disease : This attribute means whether this person has a heart disease person or not. It's numerical data.
avg glucose level : This attribute means what was the level of a person's glucose condition. It's numerical data.
Bmi : This attribute means body mass index of a person. It's numerical data.
ever married : This attribute represents a person's married status. It's categorical data.
smoking Status : This attribute means a person's smoking condition. It's categorical data.
Stroke : This attribute means a person previously had a stroke or not. It's numerical data. In this all attribute stroke is the decision class and rest of the attribute is response class.

B)Machine Learning Classifiers:

The attributes mentioned are provided as input to the different ML algorithms such as Random Forest, Decision Tree and KNN. The input dataset is split into 80% of the training dataset and the remaining 20% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analyzed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further. The different algorithms explored in this paper are listed as below.

i. Random Forest: Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

ii. Decision Tree: Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to record's attribute. On the result of comparison, the corresponding branch is followed to that value and jump is made to the next node.

iii.KNN: k-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well −

- **Lazy learning algorithm** − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm** − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

IV. RESULTS AND ANALYSIS

The results obtained by applying Random Forest, Decision Tree and KNN are shown in this section. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure. Precision (mentioned in equation (1)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (2)] defines the measure of actual positives that are correct. F-measure [mentioned in
equation (3)] tests accuracy.

Precision = (TP) / (TP +FP) (1)

Recall = (TP) / (TP+FN) (2)

F– Measure =(2 * Precision * Recall) / (Precision +Recall) (3)

• TP True positive: the patient has the disease and the test is positive.
• FP False positive: the patient does not have the disease but the test is positive.
• TN True negative: the patient does not have the disease and the test is negative.
• FN False negative: the patient has the disease but the test is negative.

In the experiment the pre-processed dataset is used to carry out the experiments and the above mentioned algorithms are explored and applied. The above mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for
different algorithms is shown below in Table 1. The accuracy score obtained for Random Forest, Decision Tree and KNN classification techniques is shown below in Table 2.

Table I VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHMS

Algorithm	True Positive	False Positive	False Negative	True Negative
KNN	839	120	4	496
Decision Tree	915	44	0	500
Random Forest	947	12	0	500

Table II ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS

Algorithm	Precision	Recall	F - measure	Accuracy
KNN	0.92	0.90	0.90	90.15%
Decision Tree	0.97	0.97	0.97	96.25%
Random Forest	0.99	0.99	0.99	99.17%

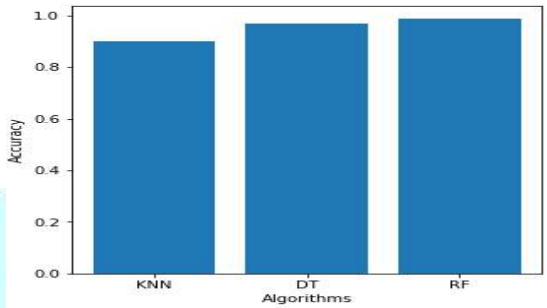


Fig 2:Bar graph representing accuracy of various algorithms used in this proposed work

V. CONCLUSION

With the increasing number of deaths due to heart stroke's, it has become mandatory to develop a system to predict heart stroke effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart stroke. This study compares the accuracy score of Random Forest, Decision Tree and KNN algorithms for predicting heart stroke using kaggle dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 99.17% for prediction of heart stroke. 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

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of paper depends largely on encouragement and guidelines of many others. We take this opportunity to express our profound gratitude to CMR Technical Campus College Management for motivating us and for providing all the facilities required for this work. We are deeply indebted to Chairman Shri C. Gopal Reddy, Secretary Smt. C. Vasantha Latha , Director Dr . A. Raji Reddy, HOD CSE Dr. K. Srujan Raju, Project Guide B.P. Deepak Kumar who always has been a constant source of inspiration for us.

REFERENCES

[1] L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi, and N. Toghianfar, "Prediction and control of heart stroke by data mining," International Journal of Preventive Medicine, vol. 4, no. Suppl 2, pp. S245–249, May 2013.

[2] S.-F. Sung, C.-Y. Hsieh, Y.-H. Kao Yang, H.-J. Lin, C.-H. Chen, Y.- W. Chen, and Y.-H. Hu, "Developing a heart stroke severity index based on administrative data was feasible using data mining techniques," Journal of Clinical Epidemiology, vol. 68, no. 11, pp. 1292–1300, Nov. 2015.

[3] M. C. Paul, S. Sarkar, M. M. Rahman, S. M. Reza, and M. S. Kaiser, "Low cost and portable patient monitoring system for e-health services in bangladesh," in 2016 International Conference on Computer Communication and Informatics (ICCCI), 2016, pp. 1–4.

[4] S. M. Reza, M. M. Rahman, M. H. Parvez, M. S. Kaiser, and S. Al Mamun, "Innovative approach in web application effort & cost estimation using functional measurement type," in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, 2015, pp. 1–7.

[5] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of heart stroke disease using machine learning algorithms," Neural Computing and Applications, vol. 32, no. 3, pp. 817–828, Feb. 2020.

[6] C.V. Krishna Veni, T. R. Shoba, On the classification of imbalanced Datasets, International Journal of Computer Science & Technology 2011; 2:145-148

Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

Sagar Yellaram

In recognition of the publication of the paper entitled

HEART STROKE PREDICTION USING MACHINE LEARNING

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 9 Issue 6 , Date of Publication:June 2021 2021-06-03 07:39:24

PAPER ID : IJCRT2106047

Registration ID : 208236

ISSN 2320-2882 IJCRT

EDITOR IN CHIEF

Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG

Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

Sumanth Kothamasu

In recognition of the publication of the paper entitled

HEART STROKE PREDICTION USING MACHINE LEARNING

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 9 Issue 6 , Date of Publication:June 2021 2021-06-03 07:39:24

PAPER ID : IJCRT2106047

Registration ID : 208236

ISSN 2320-2882 IJCRT

EDITOR IN CHIEF

Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG

Certificate of Publication

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN: 2320 - 2882

An International Open Access, Peer-reviewed, Refereed Journal

The Board of

International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

SurendharReddy Puchakayala

In recognition of the publication of the paper entitled

HEART STROKE PREDICTION USING MACHINE LEARNING

Published In IJCRT (www.ijcrt.org) & 7.97 Impact Factor by Google Scholar

Volume 9 Issue 6 , Date of Publication:June 2021 2021-06-03 07:39:24

PAPER ID : IJCRT2106047

Registration ID : 208236

EDITOR IN CHIEF

Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.97 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | IJCRT

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijcrt.org | Email id: editor@ijcrt.org | ESTD: 2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG