GUI BASED PREDICTION OF HEART STROKE STAGES BY SUPERVISED MACHINE LEARNING ALGORITHM

M.MUTHUSELVI[1],S.SINDHUMATHI[2],R.SWETHA[3],MRS.K.AMSAVALLI[4]

1# , 2# , 3#Student BE Computer science and engineering,

4# Assistant Professor, BE Computer Science and Engineering,

1# , 2# , 3# ,4#Anand Institute of Higher Technology, Chennai, India.

ABSTRACT

Heart attacks diseases are considered as the most Prevalent.Stroke Prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. The traditional prescient models or methods are as yet not powerful enough in catching the hidden information since it is unequipped for reproducing the multifaceted nature on include portrayal of the therapeutic issue space. In existing system, Vascular state is determined by measuring changes in the pressure waveforms induced through intentional variation in the device generated blood flow.To overcome this problem, predictive analytical techniques for heart stroke using machine learning model applied on given hospital dataset. The performance from the given hospital dataset with evaluation of classification report and identify the confusion matrix. Focuses on the design of a graphical user interface (GUI) for the prediction of stroke using risk parameters. Data collected from Kaggle was successfully trained and tested using Supervised machine learning algorithm.To propose a machine learning-based method to accurately predict the heart stroke by given attributes in the form of best accuracy from comparing supervise classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given healthcare department dataset with evaluation classification report, identify the confusion matrix . Result shows that the effectiveness of GUI based the proposed machine learning algorithm technique can be compared with best accuracy with precision, Recall and F1 Score.

Keyword :- *Dataset, Python, Prediction of Accuracy result.*

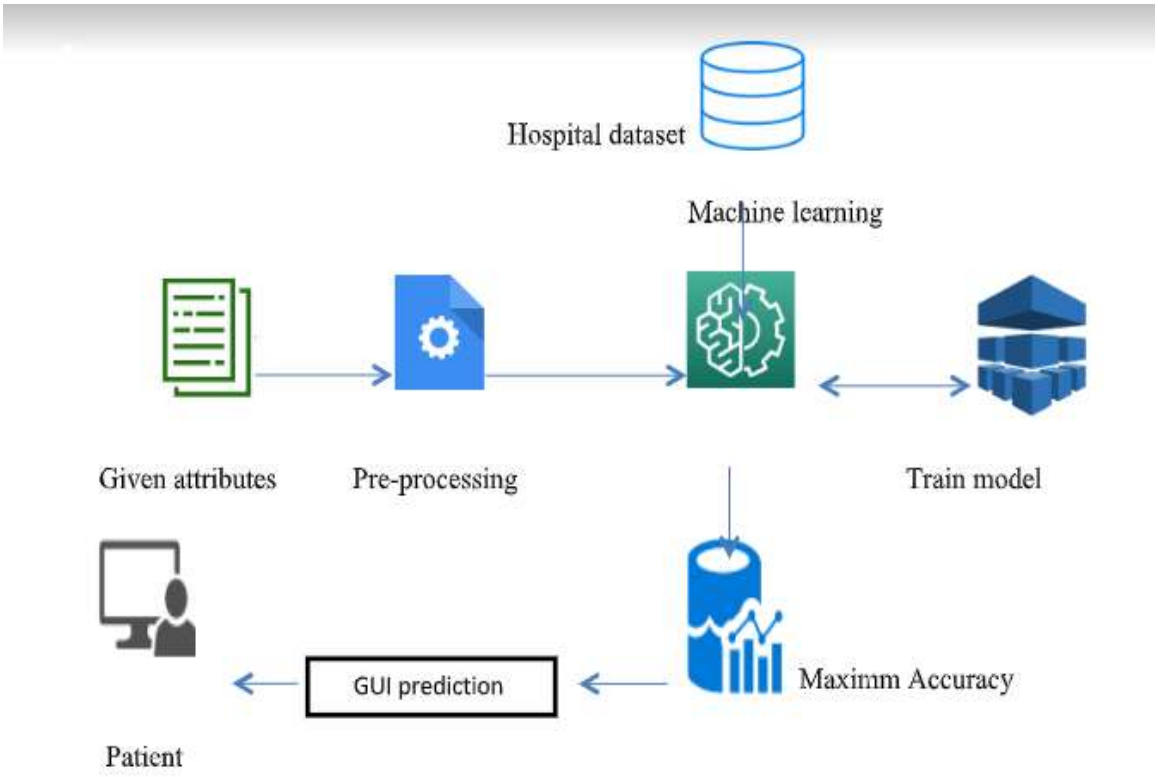
---

1. INTRODUCTION

Heart attacks diseases are considered as the most prevalent. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients . Stroke prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. Many works have been applied data mining techniques to pathological data or medical profiles for prediction of stroke. Some approaches try to do prediction on control and progression of disease. Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python.

Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and orresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance. At a high level, these different algorithms can be classified into two groups based on the way they "learn" about data to make predictions.

2. ARCHITECTURE DIAGRAM:



3. MODULES:

3.1 . DATA VALIDATION PROCESS:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset . To finding the missing value, duplicate value and description of data type whether it is float variable or integer. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time- consuming to-do list. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model and procedures .Best use of validation and test datasets when evaluating your models.

3.1.1 DATA CLEANING:

Data cleaning by rename the given dataset and drop the column etc. To analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

3.1.2 DATA PREPROCESSING:

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. To execute random forest algorithm null values have to be managed from the original raw data.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	30669	Male	3.0	0	0	No	children	Rural	95.12	18.0	NaN	0
1	30468	Male	58.0	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
2	16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	NaN	0
3	56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0

FIG 3.1.2 Before drop the given dataset

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	11788	1	48	1	0	1	2	1	3169	261	1	0
3	22348	0	60	0	0	1	2	0	1297	228	0	0
6	20811	0	42	0	0	1	2	1	2137	46	0	0
7	16165	0	65	0	1	1	3	0	10819	139	1	0
8	5841	0	22	0	0	1	2	0	2145	192	2	0

FIG 3.1.3 Data preprocessing

3.2 DATA VISUALISATION PROCESS USING SKLEARN PACKAGE:

Data visualization is an important skill in applied statistics and machine learning. Data visualization is an important skill in applied statistics and machine learning. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts . Outliers in input data can skew and mislead the training process of machine learning algorithms .

Outliers can skew the summary distribution of attribute values in descriptive statistics like mean and standard deviation and in plots such as histograms and scatterplots, compressing the body of the data. Cross-validation is a technique in which we train our model using the subset of the data-set.

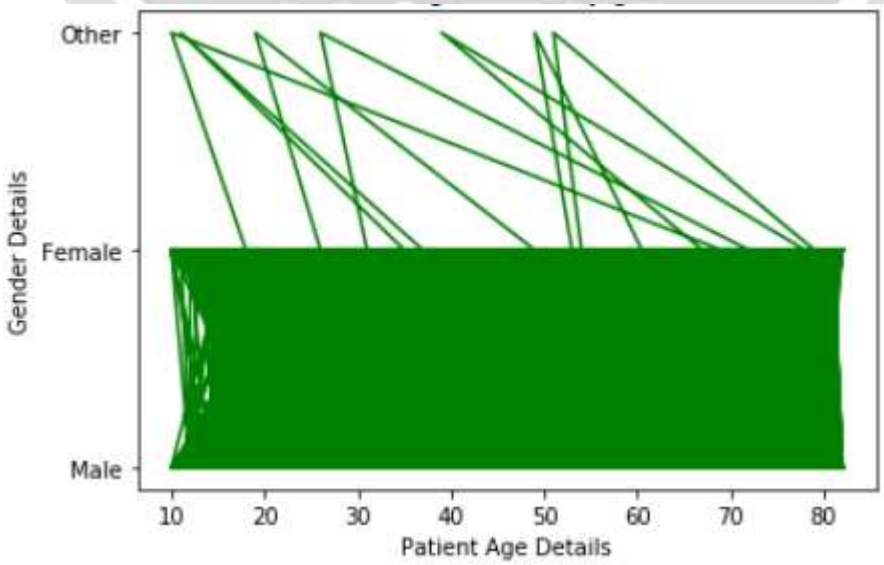


FIG 3.2. Patient Age details

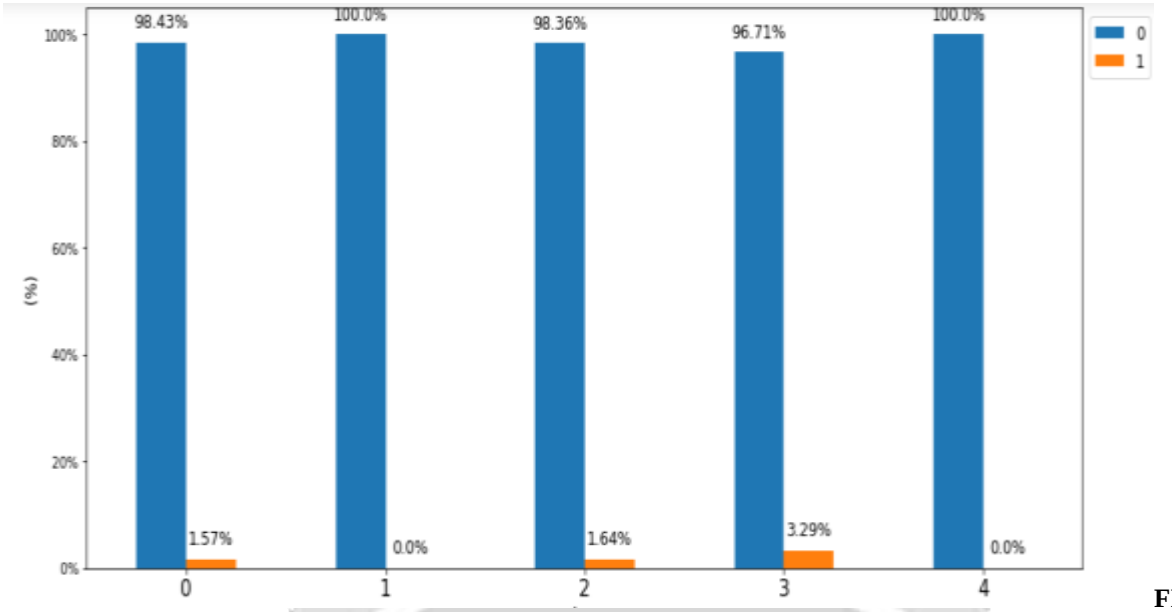


FIG 3.2 Stroke having or not by given attribute

3.3 LOGISTIC REGRESSION:

Logistic regression is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable . Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable). The independent variables should be independent of each other. Logistic regression requires quite large sample sizes.

```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      8558
           1       0.00      0.00      0.00       164

    accuracy                           0.98      8722
   macro avg       0.49      0.50      0.50      8722
weighted avg       0.96      0.98      0.97      8722

Accuracy result of Logistic Regression is: 98.10823205686769

Confusion Matrix result of Logistic Regression is:
 [[8557    1]
 [ 164    0]]

Sensitivity :  0.9998831502687544

Specificity :  0.0
```

FIG 3.3.1 Classification report of logistic regression

3.4. RANDOM FOREST:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks. Operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm can be used for both regression and classification tasks.

```
                precision    recall  f1-score   support

             0       0.98      1.00      0.99      8558
             1       0.40      0.01      0.02       164

      accuracy                           0.98      8722
     macro avg       0.69      0.51      0.51      8722
  weighted avg       0.97      0.98      0.97      8722

Accuracy result of Random Forest is: 98.10823205686769

Confusion Matrix result of Random Forest is:
 [[8555    3]
 [ 162    2]]

Sensitivity :  0.9996494508062631

Specificity :  0.012195121951219513
```

FIG 3.4.1 Classification report of random forest

3.5. SUPPORT VECTOR MACHINE:

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms.A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms.Learned SVM model representation can be used to make predictions for new data. The representation used by SVM when the model is actually stored on disk.

```
                precision    recall  f1-score   support

             0       0.98      1.00      0.99      8558
             1       0.00      0.00      0.00       164

    accuracy                           0.98      8722
   macro avg       0.49      0.50      0.50      8722
weighted avg       0.96      0.98      0.97      8722

Accuracy result of Support Vector Machines is: 98.1196973171291

Confusion Matrix result of Support Vector Machines is:
 [[8558    0]
 [ 164    0]]

Sensitivity :  1.0

Specificity :  0.0
```

FIG 3.5.1 Classfication report of support vector machine

3.6. GUI BASED PREDICTON OF HEART STROKE:

Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface.

ACCURACY CALCULATION:

FALSE POSITIVES (FP):

A person who will pay predicted as defaulter. When actual class is no and predicted class is yes.

FALSE NEGATIVES (FN):

A person who default predicted as payer. When actual class is yes but predicted class in no.

TRUE POSITIVES (TP):

A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

TRUE NEGATIVES (TN):

A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

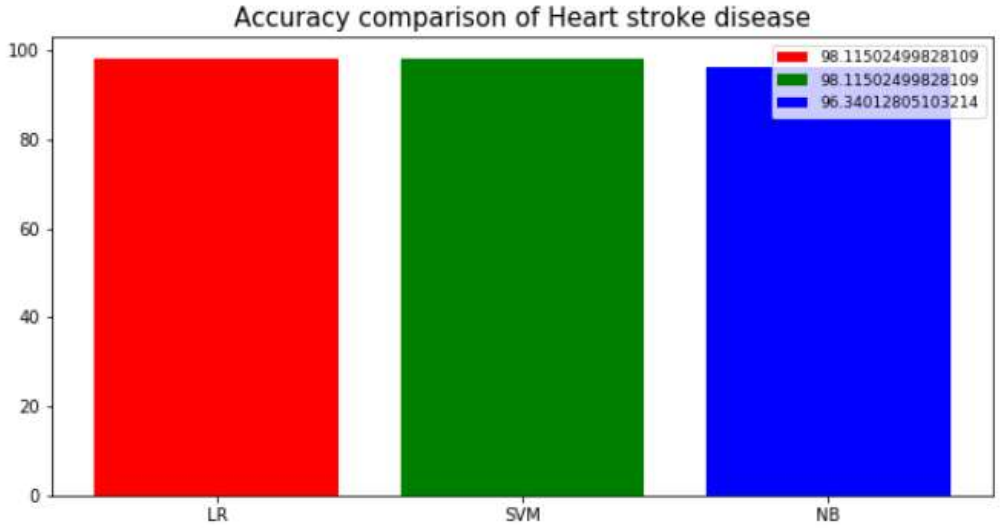
COMPARING ALGORITHM WITH PREDICTION IN THE FORM OF BEST ACCURACY RESULT:

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data .A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracy.

The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. This process is iterated throughout the whole k folds.

True Positive Rate(TPR) =TP/(TP+FN)

False Positive rate(FPR) = FP /(FP+TN)



ADVANTAGES:

The goal of this problem is to predict the status of detecting the patient having heart stroke or not by prediction accuracy results of test dataset. To reduce doctor risk in healthcare.

4. SOFTWARE DESCRIPTION:

ANACONDA NAVIGATOR:

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands.

THE JUPYTER NOTEBOOK:

The Jupyter Notebook is an open-source web application that allows to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

CONDA:

Conda is an open source, cross-platform, language-agnostic package manager and environment management system that installs, runs and updatespackages and their dependencies. The Conda package and environment manager is included in all versions of Anaconda, Miniconda, and Anaconda Repository.

NOTEBOOK DOCUMENT :

Notebook documents (or "notebooks", all lower case) are documents produced by the Jupyter Notebook App which contain both computer code (e.g. python) and rich text elements. Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc.) as well as executable documents which can be run to perform data analysis.

5.PACKAGES:

SKLEARN:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.

NUMPY:

- It is a numeric python module which provides fast math functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

PANDAS:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

MATPLOTLIB:

- Data visualization is a useful way to help with identify the patterns from given dataset.

TKINDER:

- Tkinter is the standard GUI library for Python.
- Python when combined with Tkinter provides a fast and easy way to create GUI applications.

6. RESULT:

We are using preprocessing technique to convert raw data set in to clean data set. It can be inferred from this model ,use of machine learning technique is useful in developing prediction models that can help to find the heart stroke stages and reduce the doctor's risk .

TEST DATA SET

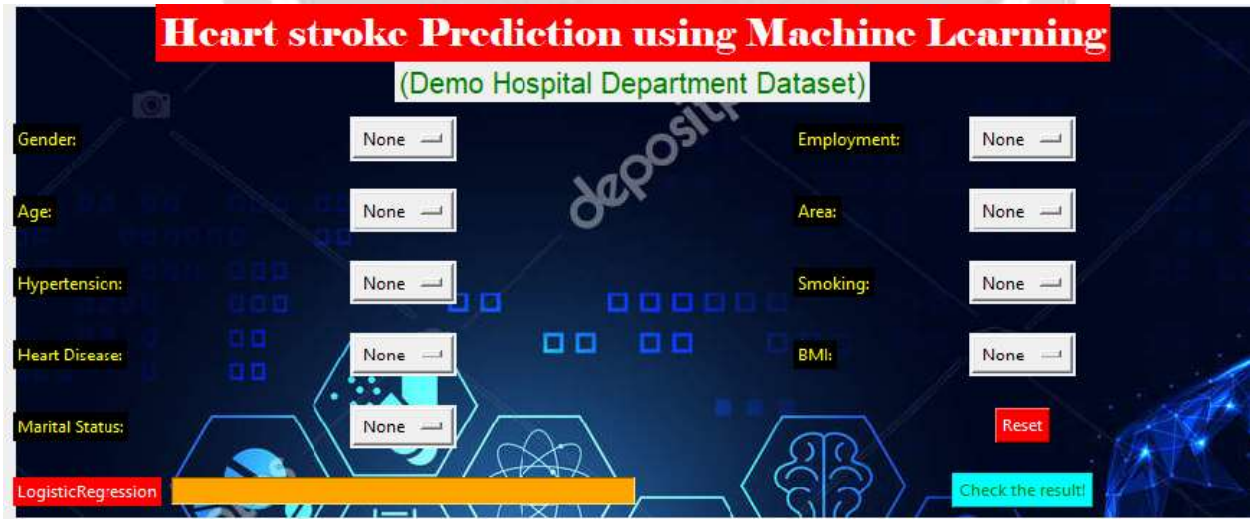
```
id,gender,age,hypertension,heart_disease,ever_married,work_type,Residence_type,avg_glucose_level,bmi,smoking_status
36306,Male,80,0,0,Yes,Private,Urban,83.84,21.1,formerly smoked
61829,Female,74,0,1,Yes,Self-employed,Rural,179.5,26,formerly smoked
14152,Female,14,0,0,No,children,Rural,95.16,21.2,
12997,Male,28,0,0,No,Private,Urban,94.76,23.4,
40801,Female,63,0,0,Yes,Govt_job,Rural,83.57,27.6,never smoked
9348,Female,66,1,0,Yes,Private,Urban,219.98,32.2,never smoked
51550,Female,49,0,0,Yes,Self-employed,Rural,74.03,25.1,
60512,Male,46,0,0,Yes,Govt_job,Urban,120.8,32.5,never smoked
31309,Female,75,0,0,Yes,Self-employed,Rural,78.71,28,never smoked
39199,Male,75,0,0,Yes,Self-employed,Urban,77.2,25.7,smokes
15160,Female,17,0,0,No,Private,Rural,78.16,21.9,
21705,Female,10,0,0,No,children,Urban,107.23,19.4,
19042,Female,47,0,0,Yes,Private,Rural,91.6,26.7,never smoked
12249,Female,42,0,0,Yes,Private,Urban,83.05,32.3,
33104,Female,67,0,0,Yes,Govt_job,Urban,236.6,24.2,never smoked
55264,Female,52,0,0,No,Self-employed,Urban,109.49,24.5,never smoked
29445,Male,73,0,0,Yes,Self-employed,Rural,109.66,40,
49013,Female,19,0,0,No,Private,Rural,88.51,22.1,
276,Male,15,0,0,No,children,Rural,101.36,22.3,
47721,Female,37,0,0,Yes,Govt_job,Urban,165.44,36.1,formerly smoked
19656,Male,59,1,0,Yes,Self-employed,Rural,101.06,33.3,formerly smoked
35806,Male,45,0,0,Yes,Private,Urban,81.54,36.3,smokes
70637,Female,44,0,0,Yes,Private,Rural,150.06,22.2,never smoked
62010,Female,13,0,0,No,children,Urban,87.79,20.5,formerly smoked
67576,Female,82,0,0,Yes,Private,Rural,205.8,36.5,formerly smoked
31774,Male,66,1,1,Yes,Private,Rural,93.79,33.1,never smoked
```

TRAIN DATA SET

```
id,gender,age,hypertension,heart_disease,ever_married,work_type,Residence_type,avg_glucose_level,bmi,smoking_status,stroke
30669,Male,3,0,0,No,children,Rural,95.12,18,,0
30468,Male,58,1,0,Yes,Private,Urban,87.96,39.2,never smoked,0
16523,Female,8,0,0,No,Private,Urban,110.89,17.6,,0
56543,Female,70,0,0,Yes,Private,Rural,69.04,35.9,formerly smoked,0
46136,Male,14,0,0,No,Never_worked,Rural,161.28,19.1,,0
32257,Female,47,0,0,Yes,Private,Urban,210.95,50.1,,0
52800,Female,52,0,0,Yes,Private,Urban,77.59,17.7,formerly smoked,0
41413,Female,75,0,1,Yes,Self-employed,Rural,243.53,27,never smoked,0
15266,Female,32,0,0,Yes,Private,Rural,77.67,32.3,smokes,0
28674,Female,74,1,0,Yes,Self-employed,Urban,205.84,54.6,never smoked,0
10460,Female,79,0,0,Yes,Govt_job,Urban,77.08,35,,0
64908,Male,79,0,1,Yes,Private,Urban,57.08,22,formerly smoked,0
63884,Female,37,0,0,Yes,Private,Rural,162.96,39.4,never smoked,0
37893,Female,37,0,0,Yes,Private,Rural,73.5,26.1,formerly smoked,0
67855,Female,40,0,0,Yes,Private,Rural,95.04,42.4,never smoked,0
25774,Male,35,0,0,No,Private,Rural,85.37,33,never smoked,0
19584,Female,20,0,0,No,Private,Urban,84.62,19.7,smokes,0
24447,Female,42,0,0,Yes,Private,Rural,82.67,22.5,never smoked,0
49589,Female,44,0,0,Yes,Govt_job,Urban,57.33,24.6,smokes,0
17986,Female,79,0,1,Yes,Self-employed,Urban,67.84,25.2,smokes,0
29217,Female,65,1,0,Yes,Private,Rural,75.7,41.8,,0
72911,Female,57,1,0,Yes,Private,Rural,129.54,60.9,smokes,0
47175,Female,49,0,0,Yes,Private,Rural,60.22,31.5,smokes,0
4057,Male,71,0,0,Yes,Private,Urban,198.21,27.3,formerly smoked,0
48588,Female,59,0,0,Yes,Private,Urban,109.82,23.7,never smoked,0
70336,Female,25,0,0,Yes,Private,Urban,60.84,24.5,never smoked,0
```

OUTPUT:

Early diagnosis of stroke is most important for the patient to reduce its impact. Aprediction model with the aid of artificial intelligence to improve over human accuracy and provide with the scope of early detection.



7. CONCLUSION:

The analytical process started from data cleaning and processing, missing value,exploratory analysis and finally model building and evaluation. Early diagnosis of stroke is most important for the patient to reduce its impact. We presented a prediction model with the aid of artificial intelligence to improve over human accuracy and provide with the scope of early detection. To compare and discuss the performance of comparative study with finding the best accuracy apply in various supervised machine learning technique from the given dataset with GUI based application by given dataset attributes.

8. FUTURE WORK:

Hospital wants to automate the detecting of the heart stroke from eligibility process (real time) based on the account detail. To automate this process by show the prediction result in web application or desktop application. To optimize the work to implement in Artificial Intelligence environment.

.

9. REFERENCE:

[1] J. S. Hochman et al., "Early Revascularization and Long-term Survival in Cardiogenic Shock Complicating Acute Myocardial Infarction," JAMA, vol. 295, no. 21, p. 2511, Jun. 2006.

[2] P. a. Heidenreich et al., "Anticipating the eventual fate of cardiovascular illness in the United States: An approach proclamation from the American Heart Association," Circulation, vol. 123, pp. 933–944, 2011.

[3] V. L. Roger, "The study of disease transmission of cardiovascular breakdown," Circ. Res., vol. 113, no. 6, pp. 646–659, 2013.

[4] D. Mozaffarian et al., Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. 2015.

[5] H. Thiele, E. M. Ohman, S. Desch, I. Eitel, and S. de Waha, "The board of cardiogenic stun," Eur. Heart J., vol. 36, no. 20, pp. 1223–1230, May 2015.

[6] L. Mill operator, "Cardiogenic Shock in Acute Myocardial Infarction the Era of Mechanical Support," J. Am. Coll. Cardiol., vol. 67, no. 16, pp. 1881–1884, 2016.

[7] A. D. Nagpal, R. K. Singal, R. C. Arora, and Y. Lamarche, "Impermanent Mechanical Circulatory Support in Cardiac Critical Care: A State of the Art Review and Algorithm for Device Selection," Can. J. Cardiol., vol. 33, no. 1, pp. 110–118, 2017.

[8] A. Reyentovich, M. H. Barghash, and J. S. Hochman, "The board of obstinate cardiogenic stun.," Nat. Fire up. Cardiol., vol. 13, no. 8, pp. 481–92, Aug. 2016.

[9] P. E. Marik, "Eulogy: aspiratory vein catheter 1970 to 2013.," Ann. Serious Care, vol. 3, no. 1, p. 38, 2013.

[10] H. J. C. Swan, W. Ganz, J. Forrester, H. Marcus, G. Precious stone, and D. Chonette, "Catheterization of the Heart in Man with Use of a FlowDirected Balloon-Tipped Catheter," N. Engl. J. Drug., vol. 283, no. 9, pp. 447–451, Aug. 1970.

[11] W. Ganz, R. Donoso, H. S. Marcus, J. S. Forrester, and H. J. Swan, "another procedure for estimation of cardiovascular yield by thermodilution in man.," Am. J. Cardiol., vol. 27, no. 4, pp. 392–6, Apr. 1971.

[12] S. N. Ahmed, F. M. Syed, and D. T. Porembka, "Echocardiographic assessment of hemodynamic parameters," Crit. Care Med., vol. 35, no. 8 SUPPL., 2007.

[13] J. A. Chirinos, "Blood vessel firmness: Basic ideas and estimation systems," J. Cardiovasc. Transl. Res., vol. 5, no. 3, pp. 243–255, 2012.

[14] T. J. Iberti, E. P. Fischer, A. B. Leibowitz, E. A. Panacek, J. H. Silverstein, and T. E. Albertson, "A multicenter investigation of doctors' information on the aspiratory corridor catheter. Aspiratory Artery Catheter Study Group," Jama., vol. 264, no. 22, p. 2928–32., 1990.

[15] K. Ikuta, Y. Wang, A. Robinson, T. Ahmad, H. M. Krumholz, and N. R. Desai, "National Trends being used and Outcomes of Pulmonary Artery Catheters Among Medicare Beneficiaries, 1999-2013," JAMA Cardiol., vol. 06520, pp. 1–6, 2017.