Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

Comparative Analysis and Implementation of Heart Stroke Prediction using Various Machine Learning Techniques

Tanisha Rakshit
School of Electronics Engineering
Vellore Institute of Technology
Vellore, India

Aayush Shrestha
School of Electronics Engineering
Vellore Institute of Technology
Vellore, India

***Abstract* – Heart disease and strokes have rapidly increased globally even at juvenile ages. Stroke prediction is a complex task requiring huge amount of data pre-processing and there is a need to automate the prediction process for the early detection of symptoms related to stroke so that it can be prevented at an early stage. In the proposed model, heart stroke prediction is performed on a dataset collected from Kaggle. The model predicts the chances a person will have stroke based on symptoms like age, gender, average glucose level, smoking status, body mass index, work type and residence type. It classifies the person's risk level by implementing various machine learning algorithms like Random Forest, Naïve bayes, Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree and Support Vector Machine (SVM). Thus, a comparative analysis is shown between the various algorithms and the most efficient one is obtained. Decision Tree algorithm was found out to be the most effective one with an accuracy of 100%.**

Keywords – Machine Learning, Data analysis, Decision Tree, SVM, KNN, Naïve Bayes

I. INTRODUCTION

Cardiovascular Diseases (CVDs) are the most common cause of death globally representing 32% of all global deaths with about 17.9 million people being affected by it. Out of these, two most common CVDs are in the form of heart attack and heart stroke accounting 85% of the total people. Heart attack is caused due to blockage of oxygen or blood supply to the heart muscle while heart stroke is caused when there is blockage of the vessel feeding the brain. Although both of the diseases are different from each other, the risk factors contributing to them are quite similar. The risk factors include unhealthy diet, tobacco use, diabetes, sedentary lifestyle, unhealthy use of alcohol, high blood pressure and family history. Detecting heart stroke and taking medical action immediately can not only prolong life but also help to prevent heart disease in the future.

Machine learning has become one of the most demanding field in modern technology. It is a form of artificial intelligence where the model can analyze the data, identify patterns and predict the outcome with minimal human intervention. Heart stroke prediction in adults can be done by using various machine learning algorithms. It has become an intrigued research problem as there are various factors or parameters that can influence the outcome. The factors include work type, gender, residence type, age, average glucose level, body mass index, smoking status of the individual and any previous heart disease.

The proposed model predicts heart stroke prediction of several individuals using various machine learning algorithms like Random Forest, K-Nearest Neighbors, Decision Tree Classifier, Support Vector Machine, Logistic Regression and Naïve Bayes based on these input factors which has been taken from the dataset on which the model has been trained.

II. LITERATURE SURVEY

In [1], heart disease prediction is done using Naïve Bayes and Genetic algorithms. The model has been trained on a UCI dataset with attributes like gender, age, resting blood pressure, cholesterol, fasting blood sugar, old peak, etc. It is a web-based machine learning application where the user inputs his medical details based on these attributes to predict his heart disease. The algorithm calculates the probability of having a heart disease and the result is displayed on the web page itself.

In [2], various classification algorithms are studied and the most accurate model is obtained for predicting the heart disease in the patient. It was found that Random Forest and XGBoost were the most efficient algorithms while K-Nearest Neighbor was found to be the most ineffective one.

In [3], a novel heart attack prediction mechanism is proposed mainly using Decision Tree Classifier algorithm. The model first learns the deep features based on the attributes provided in the dataset and
then trains on the learned features to obtain the outcome or prediction.

In [4], a survey is proposed on the various machine learning algorithms that could be used for the heart disease prediction. The authors have summarized the various algorithms and then worked towards finding the best algorithm by analysing the various features.

In [5], heart disease prediction has been performed using the four algorithms- Logistic Regression, Naïve Bayes, Random Forest and Decision Tree. The objective is to effectively study whether the patient has any heart disease. The health professional enters the input values from the patient's health report. The data is then fed into the machine learning model which provides the probability of having the heart disease.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

In [6], heart stroke prediction is analysed using various machine learning algorithms and the Receiver Operating Curve (ROC) is obtained for each algorithm. It has been implemented using Apache Spark and it shows that the Gradient Boosting Algorithm gives the highest ROC score of 0.90. The analysis of the features has been done by using univariate and multivariate plots to obtain the correlation between the several features.

III. PROPOSED MODEL

In this paper, a model is proposed to predict whether the individual will have heart stroke or not based on several input parameters like age, gender, smoking status, work type, etc. The dataset is trained on various machine learning algorithms and their performance is analysed to find out which one would be the best to effectively predict heart stroke. The accuracies obtained from each algorithm is plotted to show the comparative analysis of each algorithm. Fig 1 shows the flowchart of the proposed model. First data collection is done followed by data pre-processing to obtain a cleaned dataset with no null values or duplicate values for better training and higher accuracy. Then data visualization is performed which gives a clear idea about the dataset through the visualization graphs and makes it easier to identify the patterns, trends and outliers. After this, the dataset is split into training and testing datasets and fed into the various classification models to obtain the prediction. The confusion matrix along with the accuracies of the models are obtained to find out the most effective algorithm that could be used for the prediction. The model has also been trained using a custom input value to check for the accuracy.

A. System Flowchart

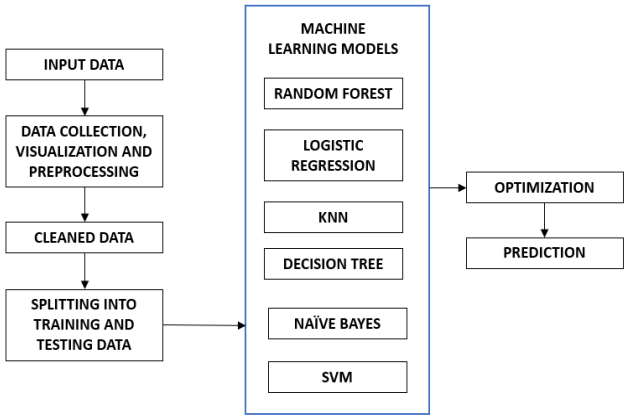

Fig 1. Flowchart of Proposed Model

B. Dataset Description

The dataset has been taken from the Kaggle website. It has 5110 rows and 12 columns. The features or attributes include id, gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, body mass index (BMI), smoking status. The label or the outcome is stroke. Excluding the ID, all the other features have been used for training the model. The independent attributes are stored in the X variable while the dependent attribute which is the stroke attribute is stored in the y variable. Fig 2 shows the dataset.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

Fig 2. Dataset for stroke prediction

C. Data Pre-processing

The dataset obtained contains 201 null values in the BMI attribute which needs to be removed. Presence of these values can degrade the accuracy of the model. Also, the categorical values are encoded into numerical values using the 'LabelBinarizer' method as training can only be done on the numerical values as it involves standardization of the attributes. Fig 3 shows the cleaned pre-processed data.

	age	gen	marital	work	residence	smoke	hypertension	heart_disease	avg_glucose_level	bmi	stroke
0	67.0	0	0	0	0	0	0	1	228.69	36.6	1
2	80.0	0	0	0	1	1	0	1	105.92	32.5	1
3	49.0	1	0	0	0	2	0	0	171.23	34.4	1
4	79.0	1	0	1	1	1	1	0	174.12	24.0	1
5	81.0	0	0	0	0	0	0	0	186.21	29.0	1
...	...	...	...	...	...	...	...	...	...	...	...
5104	13.0	1	1	3	1	3	0	0	103.08	18.6	0
5106	81.0	1	0	1	0	1	0	0	125.20	40.0	0
5107	35.0	1	0	1	1	1	0	0	82.99	30.6	0
5108	51.0	0	0	0	1	0	0	0	166.29	25.6	0
5109	44.0	1	0	2	0	3	0	0	85.28	26.2	0

4909 rows × 11 columns

Fig 3. Pre-processed dataset

D. Data Visualization

Data visualization helps to interpret the data easily through visual graphs or maps. Heatmaps are used to obtain the correlation between the attributes as shown in Fig 4. Histogram plots are used to count the frequencies of smokers and non-smokers, number of females or males, the different work types of the people as shown in Fig 5. Box plots have been used to indicate the relationship between two attributes and find out the outliers as shown in Fig 6. All these plots give important insights about the data which can later be used for the modelling. It also shows which features are more important in making the most accurate prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

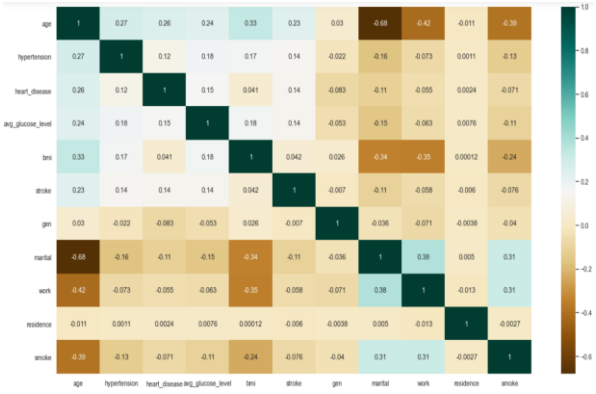

Fig 4. Heatmap to show correlation

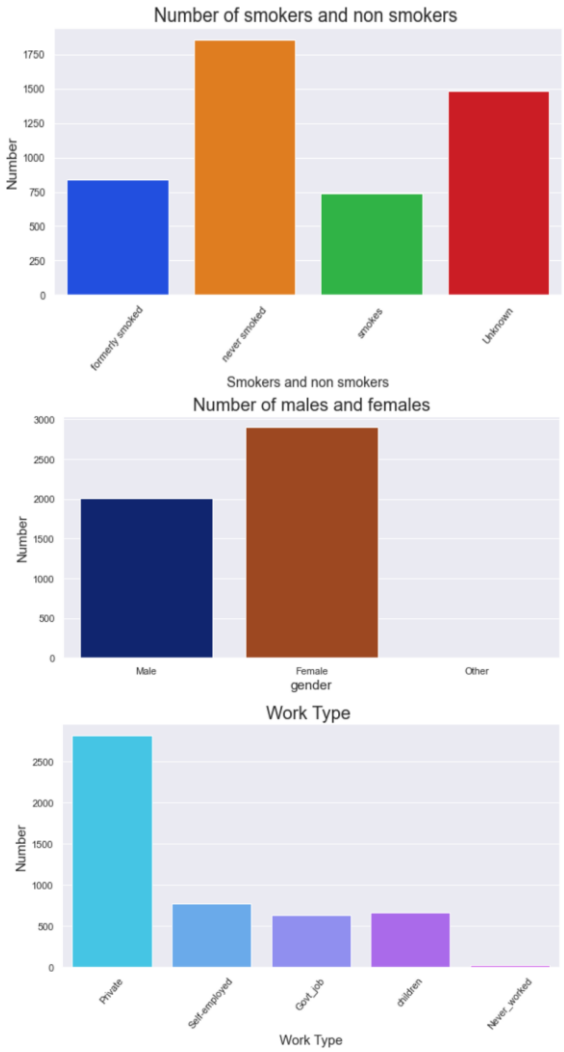

Fig 5. Histogram Plots to show frequencies

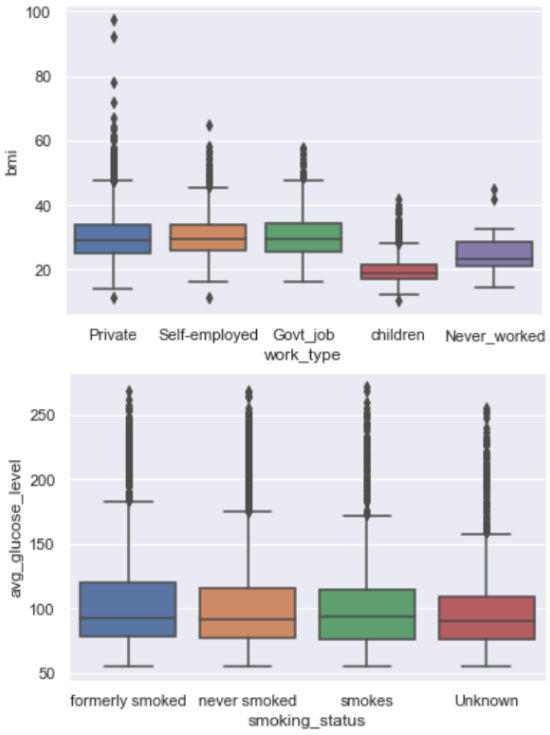

Fig 6. Box plots to show relation between 2 features

E. Data Splitting

The dataset is split into dependent and independent attributes using the train test split method of sklearn package in python. The dataset is split into 75% for training and 25% for testing. The independent features include all the input parameters like age, gender, work type, smoking status, etc while the dependent feature is stroke.

F. Classification

Random Forest- It is the most used supervised machine learning algorithm for classification and regression. It uses ensemble learning method in which predictions are based on the combined results of various individual models. Finally voting is used to find the class of the predicted value. It works by two techniques- bagging and boosting. Bagging is the process in which the entire dataset is divided into n different random subsets and each individual decision tree is created on each random subset. The trees predict on different columns and data rows and then these trees are trained to obtain the vote. Boosting is the training of individual models in a sequential way. Each individual model learns from mistakes made by previous model.

Logistic Regression- It is a method to predict a dependent variable given a set of independent variables such that dependent variable is categorical. The dependent vs independent variable is mapped to a sigmoid function. So, value of dependent variable is found either 0 or 1 for any value of independent variable. It gives the probability of occurrence of an event and several results like accuracy, ROC curve, F1 score, precision, recall, confusion matrix, etc can be obtained from logistic regression.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

K-Nearest Neighbor (KNN) – It is a simple algorithm that stores all the available cases and classifies the new data or case based on similarity measure. 'K' means the number of nearest neighbors which are voting class of new or testing data. To calculate the least distant 'k' points, mathematical equations like Euclidean distance, Manhattan distance, etc are employed. It is also called Lazy Learner because it does not have a discriminative function from the training data. It memorizes the training data and there is no learning phase of the model.

Decision Tree – It is a tree shaped diagram used to determine a course of action. Each branch of tee represents a possible decision, occurrence or reaction. It can be used for classification and regression. Classification is applied on discrete values while regression is applied on continuous values. A classification tree will determine a set of logical if-then conditions to classify problems while regression tree is used when the target value is numerical or continuous in nature. These are simple to understand, interpret and visualize.

Naïve Bayes – This algorithm is based on the Bayes theorem. The assumption taken here is that all the input features or attributes are independent of each other and it provides the conditional probability as the output based on the input parameters. Bayes' theorem calculates the posterior probability of an event (A) given some prior probability of event B represented by P(A/B) as shown in equation 1:

$$P(A|B) = (P(B|A)P(A)) / P(B) \quad (1)$$

Support Vector Machine (SVM) – It is a supervised learning method in which the model learns from the past input data and makes future prediction as output. It can be both classification and regression based. SVM works on the labelled sample data to obtain the decision boundary which produces the new unlabelled data and then the new data is plotted from which the unknown value is predicted. Distance between the support vector and the hyperplane is kept as far as possible.

IV. RESULTS AND ANALYSIS

The results obtained after applying Random Forest, Logistic Regression, KNN, Decision Tree, Naïve Bayes and SVM 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 (2)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (3)] defines the measure of actual positives that are correct. F-measure [mentioned in equation (4)] tests accuracy.

$$\text{Precision} = (TP) / (TP + FP) \quad (2)$$

$$\text{Recall} = (TP) / (TP + FN) \quad (3)$$

$$\text{F– Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

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

The above-mentioned performance metrics are obtained using the confusion matrix which is used to calculate the overall performance of the model. The confusion matrix of the Random Forest algorithm obtained by the proposed model is shown Fig 7. Table I. shows the accuracies obtained from each of the machine learning models and Fig 8 shows the comparative analysis between the different models and their accuracy scores obtained. It can be observed from the comparative graph that the best algorithm for prediction was Decision Tree with accuracy score of 100% and the worst algorithm obtained was Naïve Bayes with accuracy score of 86.840%.

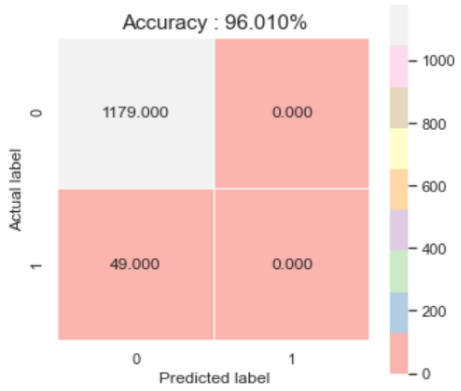

Fig 7. Confusion matrix of Random Forest

TABLE I. ACCURACIES OBTAINED USING DIFFERENT ALGORITHMS

Algorithm	Accuracies
Random Forest	**96.010%**
Logistic Regression	**95.743%**
KNN	**96.313%**
Naïve Bayes	**86.840%**
Decision Tree	**100%**
SVM	**95.743%**

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

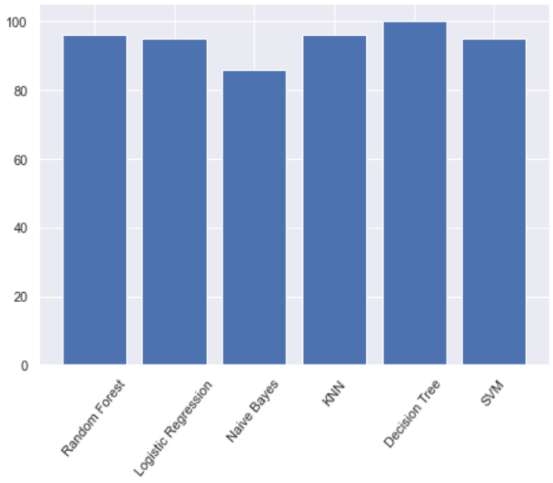

Fig 8. Comparative analysis of different models

V. CONCLUSION

As heart diseases and strokes are increasing rapidly across the world and causing deaths, it becomes necessary to develop an efficient system that would predict the heart stroke effectively before hand so that immediate medical attention can be given. In the proposed system, the most effective algorithm for stroke prediction was obtained after comparative analysis of the accuracy scores of various models. The most effective one was Decision Tree with accuracy score of 100%.

VI. FUTURE WORK

The project can be further enhanced by deploying the machine learning model obtained using a web application and a larger dataset could be used for prediction to give higher accuracy and produce better results.

VII. REFERENCES

[1] Anish Xavier, "Heart Disease Prediction using Machine Learning and Data Mining Technique", International Journal of Engineering Research & Technology (IJERT); ISSN: 2278-0181; Published by, www.ijert.org; NTASU - 2020 Conference Proceedings

[2] Pooja Anbuselvan, "Heart Disease Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT); http://www.ijert.org ISSN: 2278-0181; Vol. 9 Issue 11, November-2020

[3] Riddhi Kasabe, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT); http://www.ijert.org ISSN: 2278-0181; Vol. 9 Issue 08, August-2020

[4] Mangesh Limbitote, "A survey on Prediction Techniques of Heart Disease using Machine Learning", International Journal of Engineering Research & Technology (IJERT); http://www.ijert.org ISSN: 2278-0181; Vol. 9 Issue 06, June-2020

[5] Apurb Rajdhan, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT); ISSN: 2278-0181 http://www.ijert.org; Vol. 9 Issue 04, April-2020

[6] Maihul Rajora, "Stroke Prediction Using Machine Learning in a Distributed Environment", published in Springer, International Conference on Distributed Computing and Internet Technology; 2021; link: https://link.springer.com/chapter/10.1007/978-3-030-65621-8_15

[7] N. Komal Kumar, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers", **Published in:** 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS); IEEE Xplore

[8] Aditi Gavhane, "Prediction of Heart Disease Using Machine Learning", **Published in:** 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)