Utilizing a variety of machine learning techniques, a comparative analysis and implementation of heart attack prediction

LAKSHMANARAO BATTULA
Assistant Professor in cse dept
K L University, Vaddeswaram,
Guntur, AP, India.

USHA MATTA
Assistant Professor in cse dept
Visakha Institute of Engineering and Technology
Narava, Vizag, AP, India.

Abstract – **Even among children, heart disease and stroke rates have risen quickly worldwide. It is necessary to automate the prediction process for the early detection of symptoms associated to stroke so that it can be averted at an early stage. Stroke prediction is a complicated task involving a significant amount of data pre-processing. The suggested approach uses data obtained from Kaggle to conduct heart attack prediction. Based on symptoms such age, gender, average glucose level, smoking status, body mass index, employment type, and habitation type, the model forecasts a person's likelihood of having a stroke. It uses a variety of machine learning methods, including Random Forest, Naive Bayes, Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree, and Support Vector Machine (SVM), to categorize the individual's risk level. As a result, the several algorithms are compared, and the most effective one is found. The most efficient method, with 100% accuracy, was discovered to be the decision tree algorithm.**

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

I. INTRODUCTION

With 17.9 million persons afflicted, cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, accounting for 32% of all fatalities. Heart attacks and strokes, which account for 85% of all CVD cases, are the two most prevalent types. Heart attacks are brought on by a stoppage in the blood or oxygen supply to the heart muscle, whereas heart strokes are brought on by a blockage in the blood artery supplying the brain. Despite the fact that the two diseases are distinct from one another, they share a lot of the same risk factors. Unhealthy eating habits, cigarette use, diabetes, sedentary behavior, binge drinking, high blood pressure, and family history are among the risk factors. Heart stroke detection and prompt medical attention can both extend life and aid in preventing heart problems in the future.

One of the most difficult fields in current technology is machine learning. It is a type of artificial intelligence in which the model can examine the data, spot trends, and forecast results with little assistance from humans. Different machine learning algorithms can be used to predict heart attacks in humans. Due of the numerous variables or elements that might affect the result, it has emerged as an intriguing study subject. The variables include age, gender, type of domicile, kind of employment, average blood sugar level, BMI, smoker status, and history of heart disease. Based on these input factors, which were taken from the dataset on which the model was trained, the proposed model predicts heart stroke for a number of individuals using a variety of machine learning algorithms, including Random Forest, K-Nearest Neighbors, Decision Tree Classifier, Support Vector Machine, Logistic Regression, and Nave Bayes.

II. LITERATURE SURVEY

[1] Heart disease is predicted using genetic and naive bayes algorithms. A UCI dataset that included variables including gender, age, resting blood pressure, cholesterol, fasting blood sugar, old peak, etc. was used to train the model. The user enters his medical information into a web-based machine learning program that uses these characteristics to predict his heart condition. The algorithm determines the likelihood of getting heart disease, and the outcome is shown on the website.

[2] The most accurate model is discovered after studying a variety of categorization algorithms and forecasting the patient's heart illness. The most successful algorithms were determined to be Random Forest and XGBoost, while K-Nearest Neighbor was shown to be the least effective.

[3] Proposes a novel method for heart attack prediction that primarily makes use of the Decision Tree Classifier algorithm. The model initially trains on the learnt features to get the result or prediction, after which it learns the deep features based on the dataset's properties.

[4] A overview of the many machine learning methods that might be used to the prediction of cardiac disease is suggested .The authors described the different algorithms before attempting to choose the optimal method by examining the attributes.

[5] Uses the four methods Logistic Regression, Nave Bayes, Random Forest, and Decision Tree to predict heart disease. Effectively determining if the patient has cardiac disease is the goal. The healthcare provider inputs the data from the patient's health report. The machine learning model is then fed the data, and it calculates the likelihood that a person will develop heart disease.

[6] The Receiver Operating Curve (ROC) is obtained for each machine learning algorithm used to analyze cardiac stroke prediction. Using Apache Spark for implementation, it can be shown that the Gradient Boosting Algorithm provides the greatest ROC score, at 0.90. Univariate and multivariate plots were used for the feature analysis in order to determine the association between the various characteristics.

III. PROPOSED MODEL

A. This work proposes a model to predict the likelihood of a heart attack based on multiple input variables, such as age, gender, smoking status, kind of occupation, etc. In order to ascertain which machine learning technique would have the best chance of correctly forecasting heart attacks, the dataset is trained using many algorithms. The accuracy results from each algorithm are displayed to highlight the comparative analysis of each approach. Fig. 1 shows the flowchart of the proposed model. Following the initial phase of data collection, data preprocessing is carried out to produce a cleaned dataset devoid of null or duplicate values for improved training and increased accuracy[14].

B. The next stage is data visualization, which makes it easier to identify patterns, trends, and outliers by using visualization graphs to give a clear understanding of the dataset. After splitting the dataset into training and testing halves, the predictions are obtained by feeding the datasets into a number of classification models. The confusion matrix and model accuracies are gathered in order to determine the optimal algorithm that might be used for the prediction. The model was trained and tested using a certain input value.

C. System Flowchart

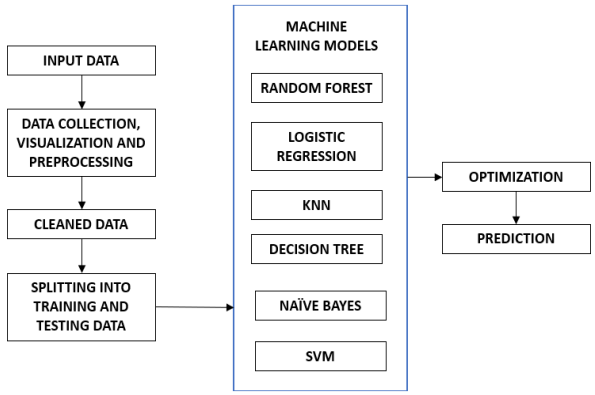

Fig 1. Flowchart of Proposed Model

D. Dataset Description

The dataset was downloaded from the Kaggle website. There are 12 columns and 5110 rows in it. The qualities or characteristics include identification, gender, age, heart disease, hypertension, previous marriages, type of work, house type, average glucose level, body mass index (BMI), and smoking status. The name or outcome is stroke. The model has been trained using all attributes except the ID. While the independent characteristics are stored in the X variable, the dependent property—the stroke attribute—is recorded in the y variable. Fig. 2 displays 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

E. Data Pre-processing

There are 201 null values for the BMI property in the retrieved dataset, which needs to be removed. Should these values exist, the accuracy of the model might be compromised. In addition, as training can only be done on numerical values because it requires standardizing the features, the 'Label Binarizer' method is utilized to encode the category values into numerical values. Fig. 3 displays the pre-processed, cleaned 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

F. Data Visualization

Data visualization uses illustrative graphs or maps to make the data easy to interpret. Heatmaps are used to determine the association between the characteristics shown in Fig. 4[12]. Histogram plots are used to count the frequencies of smokers and non-smokers, the number of males and women, and the different employment types of individuals, as Fig. 5, illustrates. Box plots have been utilized to highlight the link between two attributes and identify outliers, as seen in Fig. 6. Significant data insights are offered by each of these visualizations, which could be useful for modeling in the future[11]. It also illustrates which features are essential for generating the most accurate forecasts.

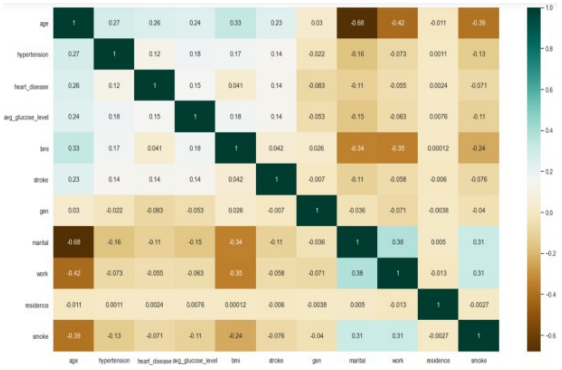

Fig 4. Heatmap to show correlation

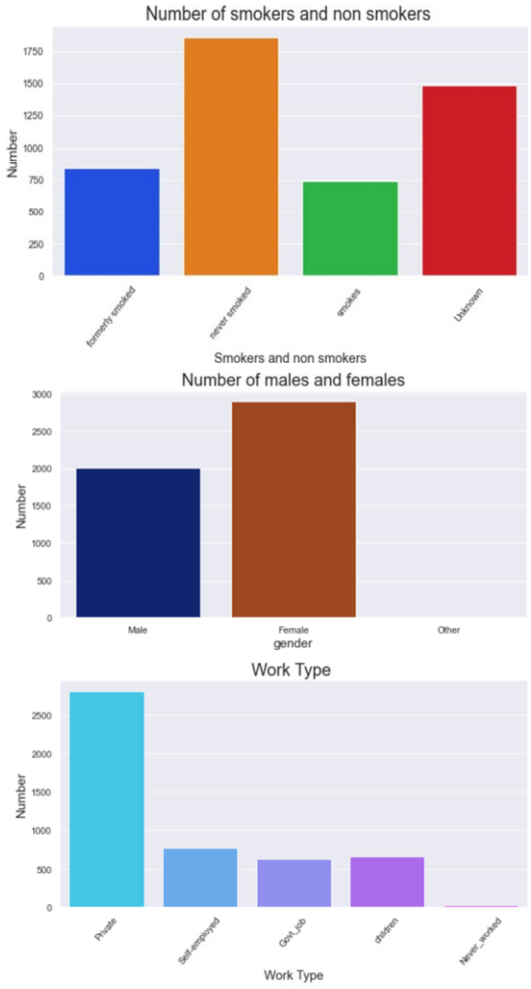

Fig 5. Histogram Plots to show frequencies

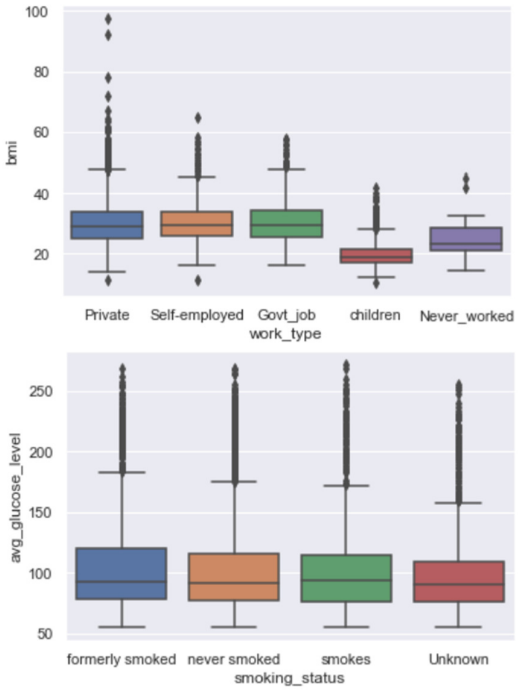

Fig 6. Box plots to show relation between 2 features

G. Data Splitting

The dataset is split into dependent and independent features using the train test split function of the Python sklearn library. Of the dataset, 75% is used for training and 25% is used for testing. Stroke is the dependent feature, while all of the input data—including age, gender, kind of employment, and smoking status—are the independent features.

H. Classification

Random Forest is the most widely used supervised machine learning technique for regression and classification. It makes use of the ensemble learning technique, in which the average output of several distinct individual models is used to make predictions. Next, voting is used to identify the class of the expected value. It uses the bagging and boosting processes to function. The entire dataset is divided into n different random subgroups using the bagging method, and a decision tree is constructed for each random subset. Before being taught to count votes, the trees make predictions based on different rows and columns of data[13]. Boosting is training individual models one after the other. Every model learns from the mistakes of the model that came before it.

Given a set of independent factors, the logistic regression approach can be used to predict a categorical dependent variable. A sigmoid function represents the relationship between the independent and dependent variables. As a result, the dependent variable's value is either 0 or 1 for every value of the independent variable. The likelihood of an event happening is provided via logistic regression, which can also yield a variety of outputs such as accuracy, ROC curve, F1 score, recall, confusion matrix, etc.

K-Nearest Neighbor or KNN It is a simple algorithm that saves all of the current instances and uses similarity measures

to classify incoming data or cases. When it comes to voting on new or experimental data, the closest neighbors are denoted by the number "K". The 'k' sites that are the closest are found using mathematical formulas such as the Manhattan distance, Euclidean distance, etc. It is sometimes referred to as the Lazy Learner since it does not have a discriminative function based on the training data. The model just memorizes the training set; there is no learning period.

A decision tree is a tree-shaped diagram that is used to select a path of action. Every branch of the tree represents a possible decision, action, or reaction. It can be used for both regression and classification. Classification is used on discrete values, whereas regression is used on continuous values. When the goal value is numerical or continuous, a regression tree is used; otherwise, a classification tree generates a set of logical if-then rules for categorizing scenarios. These are simple to understand, interpret, and visualize.

The Bayes theorem serves as the foundation for the Naive Bayes algorithm. The conditional probability is outputted depending on the input parameters under the premise that all the input characteristics or qualities are independent of one another. Equation 1 illustrates how the Bayes' theorem estimates the posterior probability of an event (A) given some prior probability of an event (B), denoted by P (A/B):

$$\mathbf{P(A \mid B) = (B \mid A)\ P(A))/P(B)} \qquad \mathbf{(1)}$$

Support Vector Machine (SVM): This supervised learning method trains the model to make future predictions by utilizing historical input data. Regression or classification may serve as its foundation. In order to identify the decision boundary, the SVM examines the labeled sample data, producing fresh unlabeled data. Next, a plot of the updated data is created to predict the unknown number. The hyper plane and support vector are separated as much as is practical. The paraphrasing tool provided by Quillwort can assist you in rapidly and effectively reworking and rephrasing your sentences.

IV. RESULTS AND ANALYSIS

This section displays the outcomes of the applications of Random Forest, Logistic Regression, KNN, Decision Tree, Naïve Bayes, and SVM. The algorithm's performance was examined using the following metrics: F-measure, Accuracy score, Precision (P), and Recall (R). The correct measure of positive analysis is provided by the precision metric (see equation (2)). The measure of true positives is defined by recall [as stated in equation (3)]. Equation (4) uses the F-measure to verify accuracy.

Precision is equal to (TP)/(TP+FP) (2)
Recall = (TP)/ (TP+FN)
Precision + Recall / (2*Precision*Recall) = F-Measure (4)

• TP True positive: the test is positive and the patient has had a stroke.
• FP False positive: the test is positive even though the patient does not have a stroke.
• TN True negative: The test is negative and the patient does not have a stroke.
• FN False negative: the test is negative even though the patient has had a stroke.

The confusion matrix, which is used to determine the model's overall performance, is utilized to derive the aforementioned performance measures. Fig. 7 displays the confusion matrix of the Random Forest algorithm produced by the suggested model. The accuracy scores from each machine learning model are displayed in Table I, and Fig. 8 presents a comparison of the models and the accuracy values they received. The comparative graph shows that Decision Tree, with an accuracy score of 100%, was the best method for prediction, while Naïve Bayes, with an accuracy score of 86.840%, was the worst.

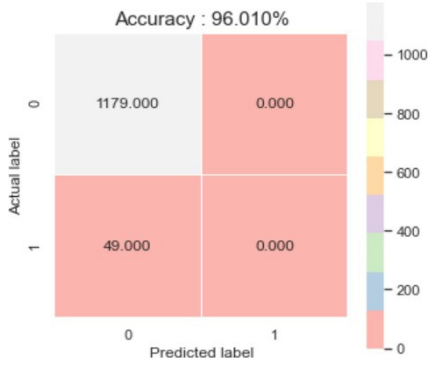

Fig 7. Confusion matrix of Random Forest

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%**

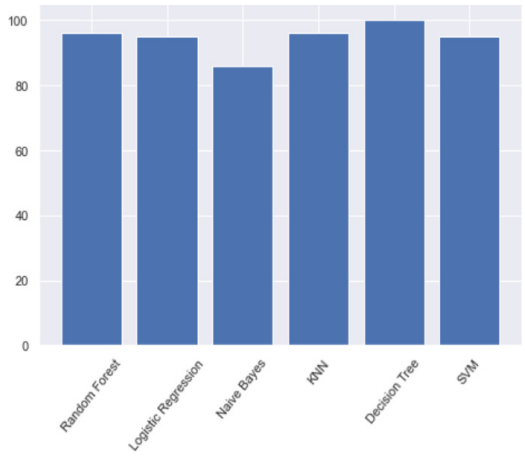

Fig 8. Comparative analysis of different models

V. CONCLUSION

It is essential to build an effective system that might detect heart attacks and strokes before they occur so that prompt medical assistance may be given. Heart illnesses and strokes are on the rise globally and are killing people. After comparing the accuracy scores of multiple models, the most efficient method for stroke prediction in the proposed system was found. With a 100% accuracy score, Decision Tree was the most efficient.

VI. FUTURE WORK

By implementing the machine learning model developed using a web application and employing a larger dataset for prediction, the project may be improved further and outcomes will be better.

REFERENCES

[1] Anish Xavier, "Heart Attack Prediction Using Machine Learning and Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), ISSN: 2078-0181, NTASU - 2021 Conference Proceedings

[2] Puja Anbuselvan, "Heart Attack Prediction Using Machine Learning Techniques", International Journal of Engineering Research and Technology (IJERT), Vol. 10 Issue 01, November-2021; available online at http://www.ijert.org ISSN: 2078-0181.

[3] Rithi Kashabe, "Heart Disease Prediction Using Machine Learning", International Journal of Engineering Research and Technology (IJERT), Vol. 9 Issue 08, August-2021; available online at http://www.ijert.org ISSN: 2078-0181.

[4] Mangish Limbitot, "A survey on Prediction Techniques of Heart Disease using Machine Learning", International Journal of Engineering Research and Technology (IJERT), Vol. 8 Issue 07, June-2027; available online at http://www.ijert.org ISSN: 2078-0181.

[5] Apoorb Rajadan, "Heart Disease Prediction Using Machine Learning", International Journal of Engineering Research and Technology (IJERT), ISSN: 2078-0181, online at www.ijert.org; Vol. 9 Issue 04, April-2021

[6] Maihool Rajoraa, "Stroke Prediction Using Machine Learning in a Distributed Environment", International Conference on Distributed Computing and Internet Technology, Springer; 2022;link: https://link.springer.com/chapter/10.1007/978-3-030-65621-8_15

[7] N. Komaal Kumaar, "Analysis and Prediction of Cardio Vascular Disease Using Machine Learning Classifiers", IEEE Xplore, 2021 6th International Conference on Advanced Computing and Communication Systems.

[8] Adithi Gavhanee, "Prediction of Heart Disease Using Machine Learning", Second International Conference on Electronics, Communication, and Aerospace Technology (ICECA), 2019.

[9] L. Aminee, 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 2014.

[10] 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.

[11] M. C. Paul, S. Sarakar, M. M. Rahaman, S. M. Reza, and M. S. Kaizer, "Low cost and portable patient monitoring system for e-health services in bangladesh," in 2016 International Conference on Computer Communication and Informatics (ICCCI), 2017, pp. 1–4.

[12] 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, 2016, pp. 1–7.

[13] P. Govindarajaan, R. K. Soundarapandiana, A. H. Gandoomi, R. Pataan, P. Jayaramana, and R. Manikandana, "Classification of heart stroke disease using machine learning algorithms," Neural Computing and Applications, vol. 32, no. 3, pp. 817–828, Feb. 2022.

[14] C.V. Krishnaa Venu, T. R. Shoba, On the classification of imbalanced Datasets, International Journal of Computer Science & Technology 2010; 2:145-148