Heart Stroke Prediction Using Machine Learning: A Comparative Analysis And Implementation

Frank Mathews Thomas
Department Of Computer Application
Amal Jyothi College Of Engineering Kanjirapally, India
Frankmathewsthomas@mca.ajce.in

Anit James
Asst.Professor Department Of Computer Application
Amal Jyothi College Of Engineering Kottayam, India
anitjames@amaljyothi.ac.in

Abstract- the primary goal of this paper is to predict coronary heart stroke. coronary heart strokes are on the upward thrust global, along with among kids and teenagers. Stroke prediction is a tough paintings that necessitates a large quantity of records pre-processing, and there's a want to automate the manner for early identity of stroke symptoms so that it may be prevented. heart stroke prediction is performed the use of a dataset inside the suggested model. primarily based on symptoms which include age, gender, average glucose degree, smoking popularity, body mass index, employment type, and residing type, the version forecasts the probability of a person having a stroke. It uses system mastering strategies which includes Random woodland, okay-Nearest Neighbor (KNN), selection Tree, to classify someone's risk level. As a result, a assessment of the various algorithms is given, and the maximum green one is determined.

Keywords – Machine Learning, KNN, Random Forest

I. INTRODUCTION

Cardiovascular diseases are one of the leading cause of mortality in the world, accounting for 32 percent of all fatalities and affecting 17.9 million peoples in the world. Out of them, the two most prevalent CVDs are heart attack and heart stroke, which account for 85 percent of all patients. A stoppage of oxygen or blood flow to the heart muscle causes a heart attack, whereas a blockage of the artery supplying the brain causes a heart stroke. Even though the diseases are distinct, the risk factors that contribute to them are very similar. Unhealthy food, cigarette use, diabetes, sedentary lifestyle, unhealthy alcohol consumption, high blood pressure, and family history are all risk factors. Detecting a heart attack and seeking medical care as soon as possible will not only help you live longer but can also help you avoid heart problems in the future. Machine learning has evolved into one of the most difficult fields in modern technology. It's a type of artificial intelligence in which a model can evaluate data, spot patterns, and anticipate outcomes with little or no human interaction. Various machine learning techniques can be used to predict heart stroke in humans. Based on these input characteristics that were obtained from the dataset on which the model was trained, the proposed model predicts heart stroke in multiple individuals using various machine learning methods such as Random Forest and K-Nearest Neighbors.

II. LITERATURE SURVEY

1. Genetic algorithms are used in the prediction of coronary heart ailment. Gender, age, resting blood pressure, ldl cholesterol, fasting blood sugar, antique peak, and different variables were used to teach the model the use of a UCI dataset. it's a web-based totally system gaining knowledge of software that uses the consumer's medical records to forecast his heart circumstance primarily based on these elements. the program evaluates the chance of developing coronary heart ailment and presentations the result on the internet page.Various classification methods are investigated, and the best accurate model for predicting the patient's heart condition is found. Random Forest was discovered to be the most efficient method, whereas KNearest Neighbor was discovered to be the least effective.
2. numerous category methods are investigated, and the pleasant correct version for predicting the affected person's heart condition is discovered. Random forest turned into observed to be the maximum efficient approach, whereas KNearest Neighbor become observed to be the least powerful.The three algorithms Random Forest, Decision Tree, and Decision Tree have been used to predict heart disease. The goal is to determine whether the patient has any type of cardiac disease. The input values from the patient's health report are entered by the health professional. The information is then incorporated into a machine learning algorithm,

DOI: 10.5281/zenodo.6364675

which calculates the likelihood of developing heart disease.

3. The Receiver Operating Curve (ROC) for each method is calculated when heart stroke prediction is examined using multiple machine learning techniques.

III.PROPOSED MODEL

A model is developed in this research to predict whether a person will have a heart attack or not depending on numerous input characteristics such as age, gender, smoking status, employment type, and so on. The dataset is trained on a variety of machine learning methods, and the results are analyzed to see which one is the most successful in predicting heart stroke. To demonstrate the comparative and analysis study of each method, the accuracies acquired from each algorithm are shown. The suggested model's flowchart is shown in Figure 1. The data is collected first, then pre-processed to provide a cleaned dataset with no null values or duplicate values for improved training and accuracy. The data is then visualized, which provides a clear picture of the dataset using visualization graphs and makes it simpler to see patterns, trends, and outliers. To achieve the prediction, the dataset is separated into training and testing datasets and fed into several categorization models. The confusion matrix, as well as the model accuracies, are obtained to choose the most effective technique for prediction. To validate for correctness, the model was also trained using a bespoke input value.

1. System Flowchart

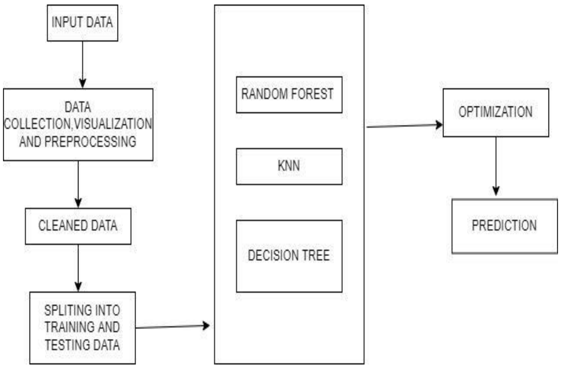
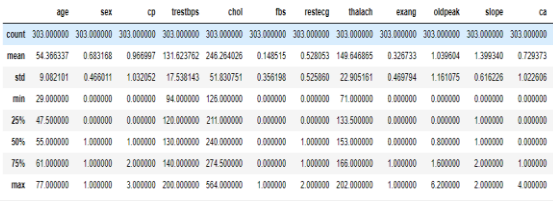

INPUT DATA
DATA COLLECTION,VISUALIZATION AND PREPROCESSING
CLEANED DATA
SPLITING INTO TRAINING AND TESTING DATA
RANDOM FOREST
KNN
DECISION TREE
OPTIMIZATION
PREDICTION

Fig 1. Flowchart Of Proposed System Design

The data was obtained through the K-platform. There are rows and columns in this table. ID, gender, age, heart disease, ever married, employment type, home type, average glucose level, body mass index (BMI), and smoking status are some of the characteristics or traits. Stroke is the label or the result. All of the other characteristics, except the ID, were utilized to train the model.

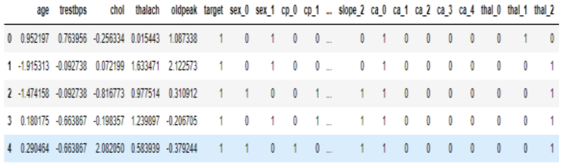


	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000

Fig 2. Dataset for Heart Stroke Prediction

2. Data Pre-Processing

The BMI property in the retrieved dataset has 201 null values, which must be deleted. The presence of these numbers can reduce the model's accuracy. Additionally, the categorical values are encoded into numerical values using the 'LIB' technique, as training can only be done on numerical values due to attribute standardisation mechanism. The data has been cleaned and pre-processed, as seen in Figure 3.



	age	trestbps	chol	thalach	oldpeak	target	sex_0	sex_1	cp_0	cp_1	...	slope_2	ca_0	ca_1	ca_2	ca_3	ca_4	thal_0	thal_1	thal_2
0	0.952197	0.763956	-0.256334	0.015443	1.087338	1	0	1	0	0	...	0	1	0	0	0	0	0	1	0
1	-1.915313	-0.092738	0.072199	1.633471	2.122573	1	0	1	0	0	...	0	1	0	0	0	0	0	0	1
2	-1.474158	-0.092738	-0.816773	0.977514	0.310912	1	1	0	0	1	...	1	1	0	0	0	0	0	0	1
3	0.180175	-0.663867	-0.198357	1.239897	-0.206705	1	0	1	0	1	...	1	1	0	0	0	0	0	0	1
4	0.290464	-0.663867	2.082050	0.583939	-0.379244	1	1	0	1	0	...	1	1	0	0	0	0	0	0	1

Fig 3. Pre-Processed Dataset

3. Data Visualisation

statistics visualisation, inside the form of visible graphs or maps, makes it easier to apprehend data. Heatmaps are used to determine the relationship among the attributes, as visible in Fig four. Histogram plots are used to matter the range of people who smoke and non-people who smoke, females and guys, and numerous work sorts of human beings, as shown in Fig five. container plots were used to highlight the connection among two variables and to discover outliers, as seen in Fig 6. All of these graphs contain important records facts that may be used inside the modelling manner later. It additionally shows which attributes are extra crucial in figuring out the maximum correct forecast.

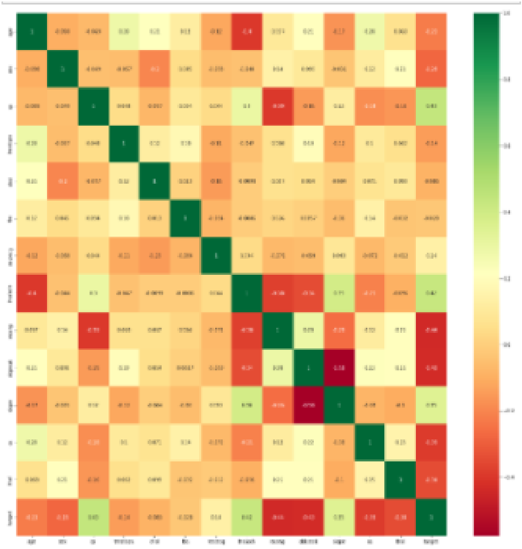


Fig 4. Heat map to show a correlation

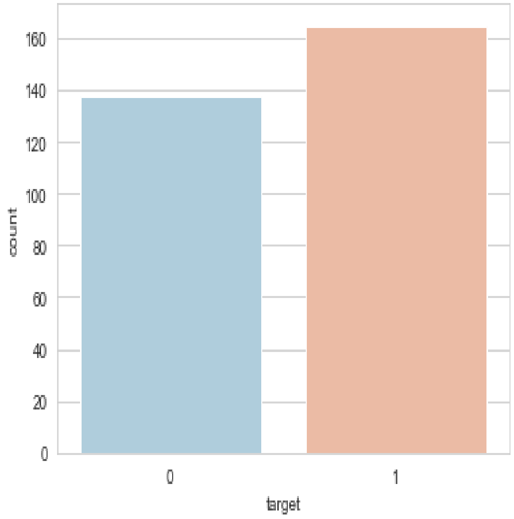


Fig 5. Number of males and females

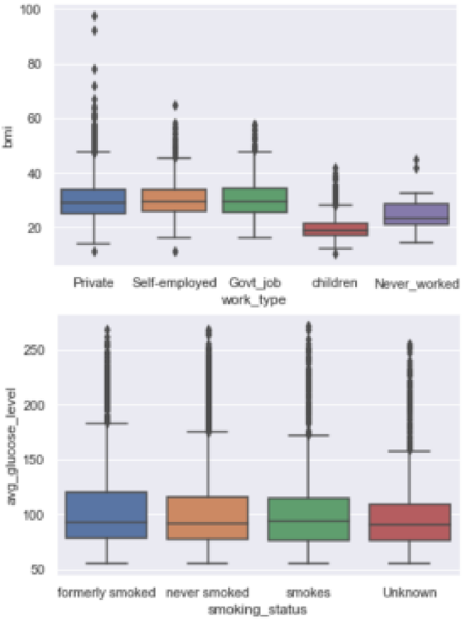


Fig 6. Box plots to show the relation between 2 features

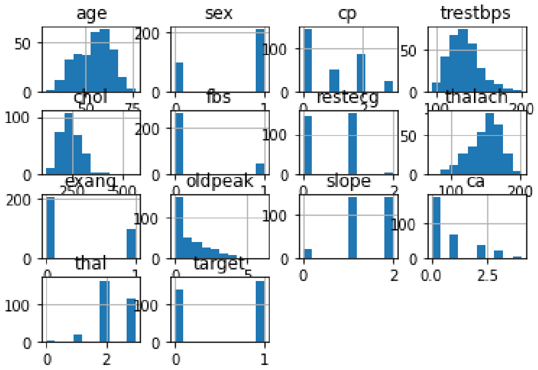


Fig 7. Histogram plots to show frequencies

4. Data Splitting

The train test split function of the package in Python is used to separate the dataset into dependent and independent characteristics. The dataset is divided into two parts: 75 percent for training and 25% for testing. All of the input characteristics, such as age, gender, employment type, smoking status, and so on, are independent features, whereas stroke is a dependent feature.

5. Classification

Random Forest- For classification and regression, it is the most widely used supervised machine learning method. It employs the ensemble learning approach, which bases predictions on the sum of the outcomes of several

DOI: 10.5281/zenodo.6364675

separate models. Finally, voting is employed to determine the anticipated value's class. It employs two methods: bagging and boosting. Bagging is the process of dividing the entire dataset into n distinct random subgroups and creating a separate decision tree for each random subset. The trees are trained to anticipate distinct columns and data rows, and then they are used to cast votes. Boosting is the process of sequentially training individual models. Each model learns from the mistakes of the preceding iteration.

K-Nearest Neighbor- It's a straightforward method that saves all of the known examples and categorizes fresh data or cases using a similarity metric. The number of nearest neighbors who are voting class of new or testing data is denoted by the letter 'K.' Mathematical formulae such as Euclidean distance, Manhattan distance, and others are used to compute the least distant 'k' locations. Because it does not have a discriminative function from the training data, it is also known as Lazy Learner. There is no learning phase for the model because it memorizes the training data.

IV.RESULTS AND ANALYSIS

We show the outcomes of Random Forest, KNN, and Decision Tree in this section. The accuracy score, Precision (P), Recall (R), and F-measure are the metrics used to assess the algorithm's performance. The precision (given in equation (2)) metric provides an accurate measure of positive analysis. Recall is a measure of true real positives [provided in equation (3)]. [Equation (4)] [Equation (3)] [Equation (4)] [Examinations of F-measure accuracy

Precision=(TP) / (TP+FP) (2)

Recall=(TP) / (TP+FN) (3)

F-Measure=(2*Precision*Recall)/ (Precision+Recall) (4)

1. TP True positive- means that the patient has had a stroke and the test has come back positive.

2. FP False-positive- the patient did not have a stroke, yet the test returns a positive result.

3. TN True negative- the patient hasn't had a stroke and the test has come back negative.

4. FN False-negative- The patient experiences a stroke, but the test comes back negative.

V. CONCLUSION

Because heart attacks and strokes are on the rise around the world and are contributing to deaths, it's more crucial than ever to develop a system that can properly predict a heart attack before it occurs, allowing for immediate medical assistance. The most successful algorithm for stroke prediction was discovered in the suggested system after analysing the accuracy ratings of several models. Decision Tree was the most accurate, with a score of 100 percent.

VI . FUTURE WORKS

The project might be improved further by implementing the machine learning model produced through a web application, and a bigger dataset could be utilized for prediction, resulting in more accuracy and better outcomes.

VII. REFERENCE

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

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 Rajdhani, "Heart Disease Prediction Using Machine Learning," International Journal of Engineering Research and Technology (IJERT); ISSN: 2278-0181 http://www.ijert.org; Vol. 9 Issue 04, April-2020.