International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

"HEART STROKE PREDICTION"

A predictive analytics approach for heart stroke prediction using machine learning and neural networks

MS. KIRUBADEVI (AP/IT)

ROOBHASRI.S PAVITHRA.S PREETHI.G

BACHELOR OF TECHNOLOGY - 1ST YEAR DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE SRI SHAKTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

(AUTONOMOUS)

COIMBATORE - 641062

ABSTRACT

Stroke is a common ailment that affects many people worldwide. Unfortunately, the statistics reveal that in the current era, about one person dies every minute due to heart stroke. This underscores the significance of utilizing data science in healthcare as it enables professionals to process vast amounts of data that are paramount in conducting research, diagnosis, treatment, and monitoring of stroke patients. In order to develop a predictive model for heart stroke. The proposed model considers various factors such as age, gender, average glucose level, smoking status, body mass index, work type and residence type to predict the likelihood of a person having a heart stroke. Data science algorithms and techniques are used to process and analyze the large amount of healthcare data available. This automated prediction process helps in identifying the risks associated with heart stroke and alerts the patient well in advance. Overall, this study highlights the significance of data science in improving healthcare outcomes and reducing the mortality rate due to heart stroke. Our research aims to develop a model that can accurately predict the risk of heart stroke in patients and classify their risk level. To achieve this goal, we will employ various data mining techniques, including machine learning algorithms like Random Forest, Naïve Bayes, Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree, and Support Vector Machine (SVM). By analysing and processing the data collected from patients, we aim to create a reliable and effective tool that can aid healthcare professionals in identifying patients at high risk of heart stroke and provide timely interventions to prevent or minimize their risk.

Keywords – Machine Learning, Data analysis, Decision Tree, SVM, KNN, Naïve Bayes Heart stroke, Random Forest

INTRODUCTION

The human heart is an essential component of the human body, and any malfunction in it could result in deprecation in other body functions. 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. The World Health Organization reports that over 10 million people worldwide die each year from heart strokes. The most effective methods of prevention are

living a healthy lifestyle and early detection. Unfortunately, the biggest obstacle in modern healthcare is providing high-quality services and accurate diagnoses. The suggested approach aims to identify these heart attacks at an early stage to prevent negative effects. The use of data mining techniques allows for the extraction of important and hidden information from the vast amount of available data. Data mining's subset of machine learning (ML) effectively manages massive, well-organized datasets. Machine learning can be used in the medical industry to diagnose, detect, and forecast a variety of diseases.

This predictive endeavour combines medical expertise with cutting-edge machine learning techniques, allowing us to sift through intricate patterns within vast datasets to identify individuals at heightened risk. By harnessing the power of artificial intelligence, we aim to empower medical professionals with timely information, enabling them to intervene and administer proactive care, ultimately sparing lives and reducing the impact of these life-altering events. In this exploration of heart stroke prediction, we delve into the amalgamation of medical knowledge and technological prowess, illustrating a promising path towards a healthier future.

The field of machine learning is increasingly in demand in modern technology. It is a form of artificial intelligence where the model can analyse the data, identify patterns and predict the outcome with minimal human intervention. Identifying the possibility of heart stroke in adults is a complex issue that can be addressed by utilizing different machine learning techniques. The research on this topic has gained significant interest due to the involvement of multiple parameters that can affect the outcome. These parameters may comprise factors such as occupation, gender, type of accommodation, age, average glucose level, body mass index, smoking habits, and any prior history of heart illness.

The suggested approach aims to identify these heart attacks at an early stage to prevent negative effects. The use of data mining techniques allows for the extraction of important and hidden information from the vast amount of available data. Data mining's subset of machine learning (ML) effectively manages massive, well-organized datasets. Machine learning can be used in the medical industry to diagnose, detect, and forecast a variety of diseases. The major objective of this research is to give medical professionals a tool for early heart stroke detection. As a result, patients will receive effective care and serious repercussions will be avoided. To uncover hidden discrete patterns and analyses the provided data, ML plays a crucial role. After data analysis, machine learning approaches aid in the early detection and prediction of heart attacks.

The model presented in this study aims to predict the risk of heart stroke for individuals based on various input factors, using a range of machine learning algorithms such as Random Forest, K-Nearest Neighbors, Decision Tree Classifier, Support Vector Machine, Logistic Regression, and Naïve Bayes. These algorithms have been trained on a specific dataset, allowing them to detect and analyse hidden patterns within the data. Machine learning techniques are essential in revealing insights from complex datasets and providing accurate predictions of health conditions such as heart stroke. 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

LITERATURE SURVEY

Data mining and machine learning techniques have shown great efficacy in various healthcare applications, particularly in medical cardiology. The growing amount of medical data provides researchers with a unique opportunity to develop and test new algorithms, especially in the identification of risk factors and early signs of heart disease, which remains a leading cause of mortality in developing countries. Naïve Bayes and Genetic Algorithms are commonly used for heart disease prediction, trained on a dataset with attributes such as age, gender, resting blood pressure, cholesterol, fasting blood sugar, and old peak. Web-based machine learning applications allow users to input their medical details to predict heart disease. Additionally, early detection of heart stroke through analysis of certain attributes has become possible with the advancement of technology in the medical field resulting in accurate diagnosis and analysis which can lead to timely action and life-saving measures. This study aimed to predict the occurrence of strokes in patients using machine learning algorithms. A dataset was obtained from Kaggle, and nine algorithms were employed, including Linear Discriminant Analysis, Logistic Regression, Gaussian Naive Bayes, Support Vector Machine, K-Nearest Neighbor Classifier, Random Forest Classifier, Bagging Classifier, Ada Boost Classifier, and Gradient Boosting Classifier. The attributes provided in the dataset were analysed to observe patterns and accurately predict the occurrence of strokes. The data were divided into training and testing datasets for further analysis. The Random Forest algorithm was found to be the most accurate with a 95.10% accuracy rate. Another research study used a different approach to predict strokes on the CHS dataset, utilizing the decision tree algorithm for feature extraction and principal component analysis.

They used a neural network classification algorithm to construct the model they got 97% accuracy. Chin et al 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.

A Research conducted by Shah et al. (2020) aimed to construct a machine learning model for predicting cardiovascular disease. To accomplish this, the team used data from the Cleveland heart disease dataset from the UCI machine learning repository that included 303 instances and 17 attributes. Various supervised classification techniques, including naive Bayes, decision tree, random forest, and k-nearest neighbor (KKN), were implemented. Results showed that the KKN model demonstrated the highest level of accuracy at 90.8%, highlighting the potential usefulness of machine learning algorithms for predicting cardiovascular disease. The study also emphasized the importance of choosing the most appropriate models and techniques to attain optimal outcomes. Additionally, four algorithms - Logistic Regression, Naive Bayes, Random Forest, and Decision Tree - were also implemented to predict heart disease, focusing on efficiently detecting the presence of heart disease in patients. 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.

Numerous studies have employed machine learning techniques to forecast the occurrence of heart strokes. One such study was carried out by Govindarajan et al, where they utilized a blend of text mining and machine learning algorithms to classify heart stroke diseases. The researchers collected data from 507 patients for their study. 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%. In their study, Amini and colleagues [4], [12] aimed to forecast the occurrence of stroke. They collected data from 807 participants, both healthy and unhealthy, and categorized 50 potential risk factors including diabetes, smoking, hyperlipidaemia, cardiovascular disease, and alcohol use. The researchers utilized c4.5 decision tree algorithm and K-nearest neighbor techniques to predict stroke incidence. The accuracy rate for the former was 95%, while the latter showed an accuracy of 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.

PROPOSED SYSTEM

In this study, a model is created to predict whether a person would experience a heart attack based on a variety of input characteristics, including age, gender, smoking status, type of employment, etc. The collection is skilled in numerous machine learning techniques, and the outcomes are examined to determine which one is most effective for forecasting cardiac attacks. To demonstrate the comparative and analysis study of each method, the accuracies acquired from each algorithm are shown. The model's flowchart is shown in Figure 1. To increase training and accuracy, the data is first gathered, after which it is pre-processed to produce a cleaned dataset without null or duplicate values. The data is then displayed, which uses visualization graphs to provide a clear image of the dataset and makes it easier to see patterns, trends, and outliers. The dataset is split into training and testing datasets and fed into several categorization models in order to achieve the prediction.

1. System Flowchart

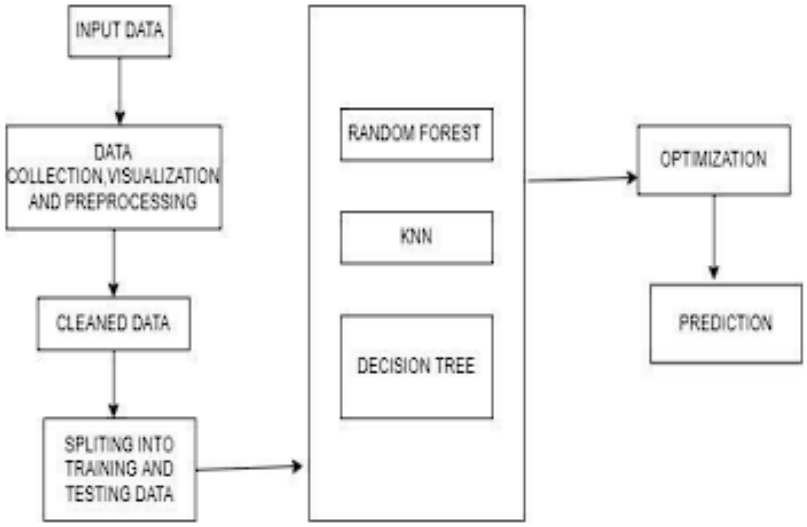


Fig 1. Flowchart Of Proposed System Design

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 data is fed into model which predicts the probability of having heart stroke. Figure 2 shows the entire process involved.

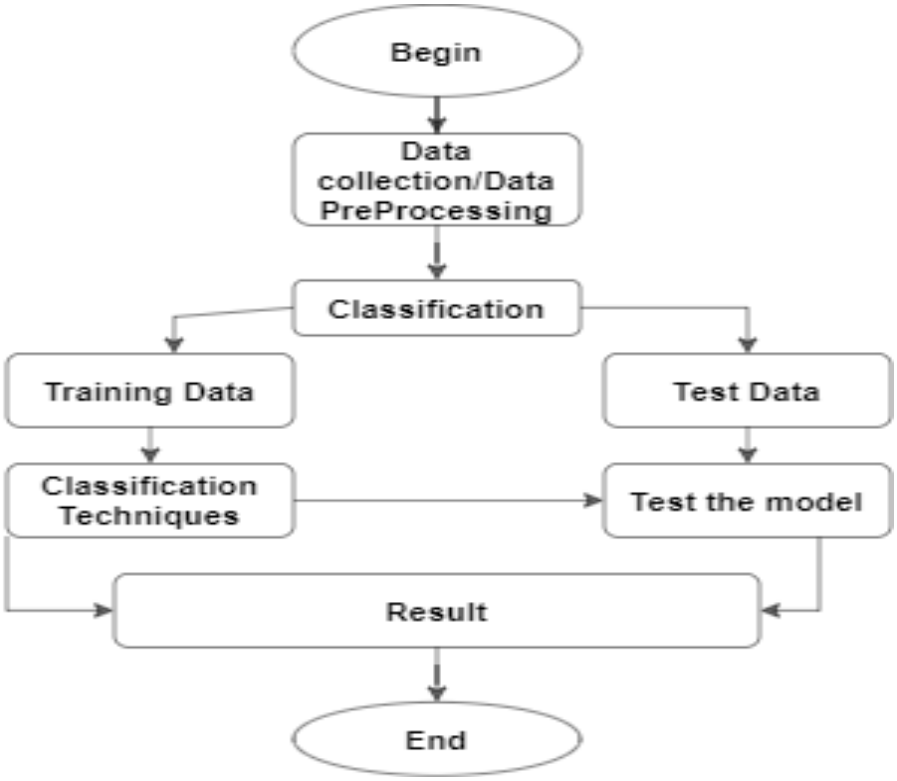
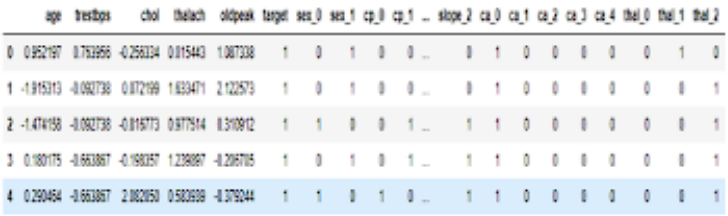


Fig 2: Generic Model Predicting Heart Stroke

2. Data Pre-Processing

The obtained dataset's BMI property has 201 null values, all of which need to be removed. The accuracy of the model may be hampered by the existence of these numbers. The 'LlB' technique is used to encode the categorical values into numerical values because training can only be done on numerical values due to the attribute standardization mechanism. Figure 3 illustrates how the data has been pre-processed and cleaned.



	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

It is simpler to understand statistics when they are visualized as clear graphs or maps. As shown in Fig. 4, heatmaps are utilized to ascertain the link between the characteristics. The distribution of smokers and non-smokers, females and males, and various work types of persons are measured using histogram plots, as illustrated in Fig. 5. As can be seen in Fig 6, container plots were used to draw attention to the relationship between two variables and to identify outliers. Each of these graphs contains crucial data points that can later be used in the modelling process. Additionally, it demonstrates which characteristics are more significant in determining the most accurate forecast.

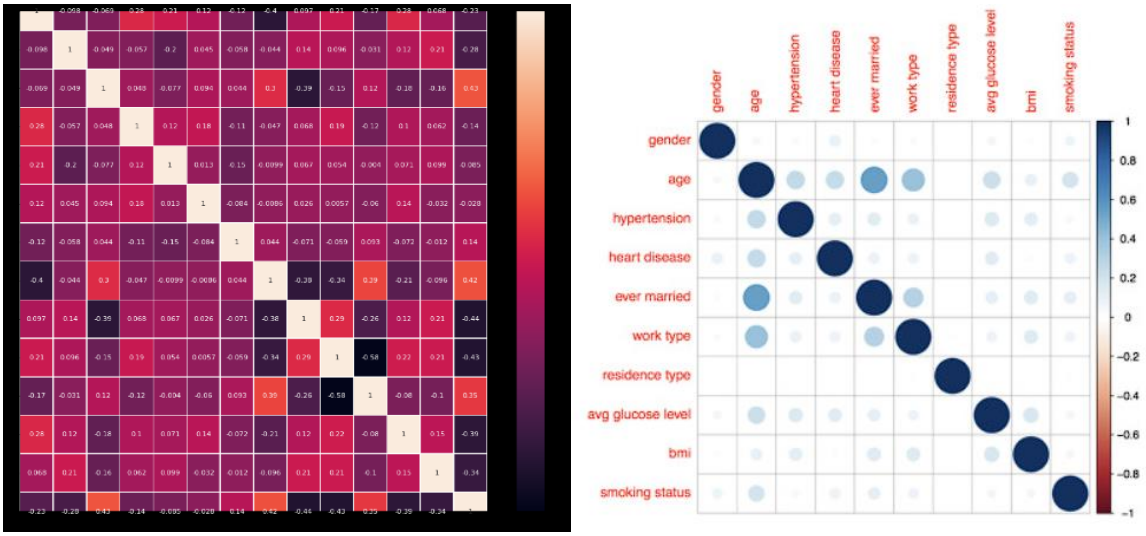


Fig 4. Heat map to show a correlation

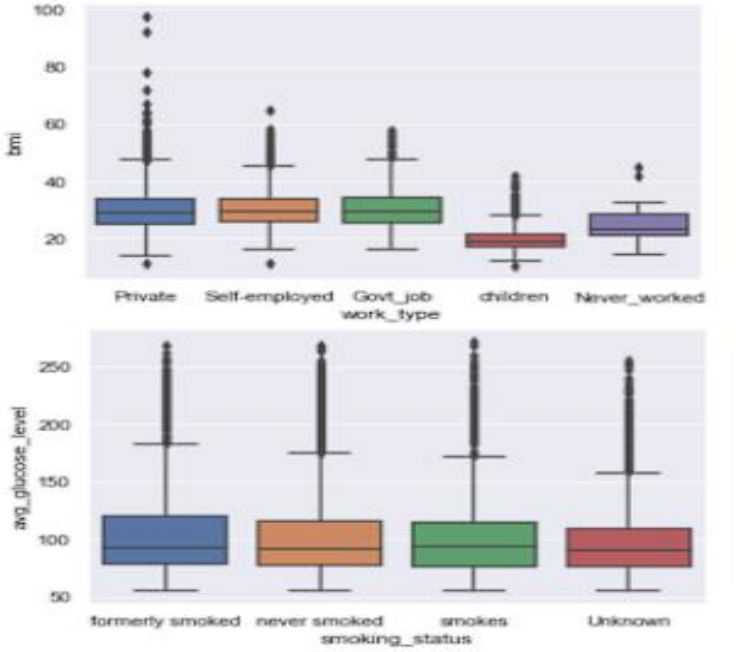


Fig 5. Box plots to show the relation between 2 features

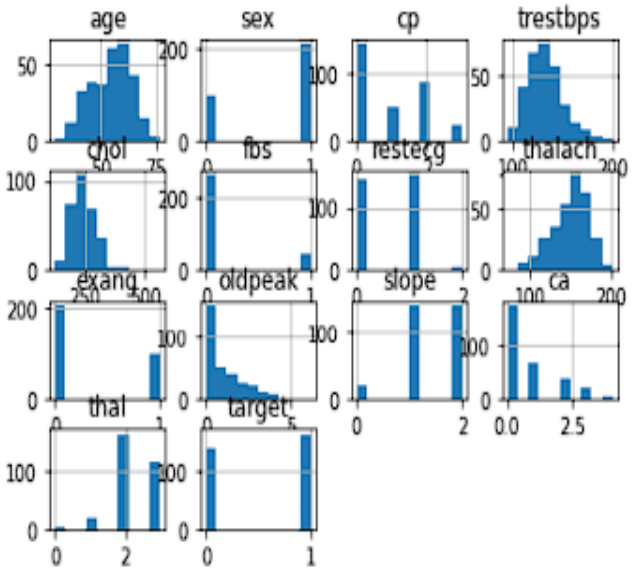


Fig 7. Histogram plots to show frequencies

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.

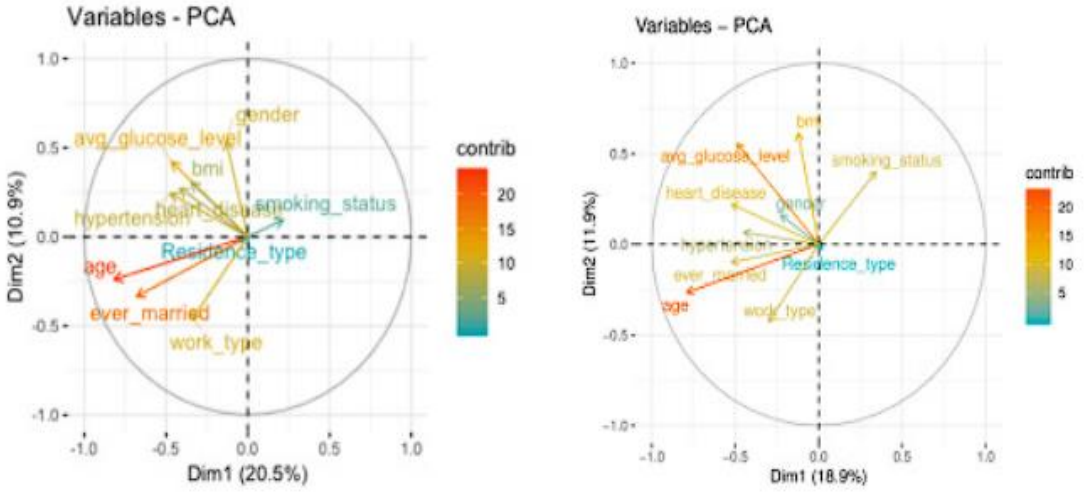
5. Classification

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 separate models. Finally, voting is employed to determine the anticipated value's class. It employs two methods: bagging and boosting. Bagging and boosting are two distinct machine learning techniques for creating decision trees. Bagging involves dividing the dataset into separate, randomized subsets and creating a decision tree for each subgroup. These trees then use their learned data to cast votes. Boosting, on the other hand, involves training individual models sequentially, with each model learning from the mistakes of its predecessors. K-Nearest Neighbor is a simple approach that saves known examples and uses them to categorize new data using a similarity metric. The number of nearest neighbors that vote for a new data's class is denoted by 'K.' Unlike other methods, it does not have a discriminative function from the training data and is therefore known as a Lazy Learner. Mathematical formulae such as Euclidean distance and Manhattan distance are used to compute the least distant K locations. There is no learning phase for the model because it memorizes the training data.

Relation between principal components with patient attributes

(a) Illustrates that the contribution of patient's residence type is minimum to the two principal components. We have noticed that there is a correlation between the age and marital status of individuals, and these variables significantly impact the first principal component. Additionally, we have found that smoking status and average glucose levels are independent of each other, implying that they offer unique insights into the feature space. However, smoking status and age point opposite to each other, indicating that they provide similar information, but have a diverging characteristic.

(b) We found that there is a significant correlation between the average glucose level and the presence of heart disease. Additionally, the age variable appears to have the greatest influence on the first two principal components. Interestingly, the feature vectors in the two-dimensional feature space align with the distribution of the imbalanced dataset.



(a) (b)

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 analysed 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:

Regression and classification both employ Random Forest methods. The data is organized into a tree, and predictions are based on that tree. Even with a substantial number of record values missing, the Random Forest algorithm can still produce the same results when applied to huge datasets. The decision tree's generated samples can be preserved and used to different sets of data.

In random forest, there are two stages: first, generate a random forest, and then, using a classifier produced in the first stage, make a prediction.

ii. Decision Tree:

The central node of the Decision Tree algorithm represents the dataset properties, and the outside branches provide the result. Decision trees are used because they are quick, dependable, simple to understand, and require very little data preparation. In a decision tree, the class label prediction comes from the tree's root. The root attribute's value is contrasted with the record's attribute. The matching branch is followed to the value indicated by the comparison result, and a jump is then made to the following node.

iii. KNN:

The supervised machine learning (ML) technique k-nearest neighbors (KNN) can be applied to classification and regression predicting issues. However, it is primarily employed in industry for classification and forecasting problems. The next two characteristics would accurately describe KNN:

- Lazy learning algorithm −KNN is a lazy learning algorithm since it uses all of the data for training while classifying rather than having a separate training phase.
- Non-parametric learning algorithm −KNN is another example of a non-parametric learning algorithm because it makes no assumptions about the underlying data.

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 metric provides an accurate measure of positive analysis. Recall is a measure of true real positives. Examinations of F-measure accuracy.

(1) Precision=(TP) / (TP+FP)
(2) Recall=(TP) / (TP+FN) (3) F-Measure=(2*Precision*Recall)/ (Precision + Recall).

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.

FUTURE WORKS

Thus, in future, we plan to integrate the electronic records dataset with background knowledge on different diseases and drugs using Semantic Web technologies. One potential approach for achieving interoperable publication of electronic health records to the research community is through the use of knowledge graph technologies. Incorporating additional background knowledge from other datasets may also help improve the accuracy of stroke prediction models. Our next steps include collecting our institutional dataset for benchmarking machine learning methods for stroke prediction and performing external validation of our proposed approach. 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.

CONCLUSION

It is essential to create a system that can anticipate heart attacks precisely and effectively given the rise in heart stroke-related fatalities. Finding the best effective ML algorithm for heart stroke diagnosis was the study's driving force. This study uses the Kaggle dataset to examine the accuracy scores of the Random Forest, Decision Tree, and KNN algorithms for predicting heart attacks. The outcome of this study shows that the Random Forest algorithm, which has an accuracy score of 99.17% for heart attack prediction, is the most effective algorithm. The work can be improved in the future by creating a web application based on the Random Forest method and by utilizing a larger dataset compared to the one utilized in this analysis, which was only a small sample.

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. Goel R Heart Disease Prediction Using Various Algorithms of Machine Learning, http://dx.doi.org/10.2139/ssrn.3884968.

4. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011; 3:67

5. Sivapalan G., Nundy K., Dev S., Cardiff B., Deepu J. ANNet: a lightweight neural network for ECG anomaly detection in IoT edge sensors. IEEE Transactions on Biomedical Circuits and Systems (2) (2022)

6. R.S. Jeena, S. Kumar, Stroke prediction using SVM, in: Proc. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2016, pp. 600–602, http://dx.doi.org/10.1109/ICCICCT. 2016.7988020.

7. Koh H.C., Tan G., *et aal. Data*mining applications in healthcare J. Health. Inf. Manage., 19 (2) (2011), p. 65