PROJECT REPORT – STROKE PREDICTION USING DECISION TREE

Shreeya Sudhesh Rao
200953176

Introduction

Stroke is a dangerous disease that occurs in a person which can cause brain damage or death. Stroke can be either an ischemic stroke or a haemorrhagic stroke. Ischemic stroke is the type of stroke that occurs when blood vessels to the brain are blocked by particles or clots. Haemorrhagic stroke occurs when blood artery ruptures in the brain. This is a disease of great concern as it is one of the major reasons of death, especially among the senior citizens. Stroke damages the lives of around 15 million people all around the world every year. However, this disease can be successfully avoided by early recognition and by making lifestyle modifications like a better diet, BMI, glucose levels and by avoiding unhealthy habits. Stroke recognition has been made possible with the advancements in science and technology. Accurate predictions can be made with Machine Learning models. The main aim of this report is to present a Decision Tree based prediction model which efficiently predicts the probability of a person susceptible to stroke, by analysing his information like age, glucose levels, BMI and habits.

Literature survey

1. Amini, L et al. [1] explains the steps to investigate 50 risk factors for stroke. For collecting data, data pre-processing, and data mining, K Means Algorithm, andthe C4.5 algorithm was used. The three phases in the entire process were the learning phase, the test phase, and the analysis phase. The learning phase consists of a model created from training data that describes the different classes and labels that are going to be present.In the test phase, the accuracy of the classification model is determined using the test data set. For a record with unspecified values, two methods were used to deal with it and they are, the use of average property values and using mean values. Analysis using C4.5 and K means algorithm using Java was then performed. For performance analysis, the sensitivity, specificity, accuracy, and precision weredetermined and plotted. The accuracy of the K Means algorithm and WEKA was 95.42% and 94.18% respectively. The C4.5 algorithm outruns the K Means algorithm in terms ofthese factors and was hence chosen as the algorithm for data analysis. Theadvantage of the paper is it is easy and simple to understand and implement.The drawback is there was insufficient information on the algorithms used.

2. M. Sheetal Singh et al. [2] discusses an AI system that takes in a patient's medical details and predicts the possibility ofa stroke. The CHS dataset is used for this purpose. Data collection, Data pre-processing, Feature Extraction, Dimension Reduction, Classification, and result analysis are the different steps used in this process. For data pre-processing work, Removal of duplicate, erroneous, noisy, and inconsistent data. Thedecision tree is used for the feature selection process.The principal component analysis is then used for dimensionality reduction. In the classification work three algorithms- Decision Tree, Naïve Bayes, and Neural Networks were used for classification but the neural network was the best fit. The back propagationneural network classification algorithm for the classification Model has a 97.7% accuracy. The advantage here is that Neural Networks was clearly elaborated Disadvantage is a presentation in the form of plots.

3. R S Jeena et al. [3] aims to develop a stroke prediction model using a Support VectorMachine by evaluating the performance of different support vector machinemodels with various kernels. Dataset was collected from the International Stroke Trial Database. SupportVector Machines is a model widely used for classification purposes. The concept here is that given a set of inputs with their corresponding class labels, we have to be able to predict whether a person has a stroke or not. The first step is data preprocessing. Processing was done to remove inconsistent data and after pre-processing, 350 samples were chosen for further processes. SVM was implemented in MATLAB and different kernels were used on the data and out of the lot, the linear kernel proved to be the best choice. The accuracy of the model was 91% and different factors like Sensitivity, Specificity, Precision, F1 score, and Accuracy were determined. The ability of the method to identify stroke-affected ones is given by Sensitivity. Specificity measures the ability of the technique to identify non- stroke cases. The fraction of correct classifications to the total number of classifications is given by accuracy. Precision is the likelihood that a retrievedcase is relevant. The harmonic mean of precision and recall (sensitivity) givesan F1 Score. The advantages of this paper are a clear explanation and description of the technologies used. The disadvantage is that data pre- processing was not explained elaborately

4. M. Mahmud et al. [4] proposes an early prediction of stroke diseases by using different machine learning approaches. Informational collection utilized hasbeen acquired from the medical clinic of Bangladesh. Ten different classifiers have been trained: they are, Logistics Regression, Stochastic Gradient Descent, Decision Tree Classifier, AdaBoost Classifier, Gaussian Classifier, Quadratic Discriminant Analysis, Multi-layer Perceptron Classifier, KNeighbors Classifier, Gradient Boosting Classifier, and XGBoost Classifier for predicting thestroke. In the paper, Python and Sci-kit learn libraries have been utilized. This paper is divided into three parts, these are Data description, machine learning classifiers & evaluation matrices, and implementation procedures. The Data pre- processing step checks the data for missing values and replaces them with the mean/median value ofthatfeature.Thenwenormalizethedataandlabelandencode it to categoricalvalues.Following this step,we performthesplittingofdatafortrainingandtesting.Thenextstepistotrainthe algorithm using the model. Here, the weighted voting classifier is implemented to increase the accuracy Confusion matrix is created. The advantage of this paper is exposure to multiple algorithms and the drawback is that the weightedvoting classifier is complex

5. Gangavarapu Sailasya et al. [5] uses 6 machine learning models for the prediction of stroke.Data Set was initially cleaned and made available for the model to use.Label Encoding is initially performed and data is split into training and test data sets. After training the model and getting a satisfactory accuracy, an HTML and flask app is developed. The first step of procuring a dataset is done with the help of the Kaggle website. After data pre-processing, Logistic Regression, Decision Tree Classification algorithm, Random Forest Classification algorithm, K- Nearest Neighbour algorithm, Support Vector Classification, and NaïveBayes Classification algorithm are used. Out of all the algorithms chosen, Naïve Bayes Classification performs best with an accuracy of 82%. After building six different models, they are compared using five accuracy metrics namely Accuracy Score, Precision Score, Recall Score,F1Score, and Receiver Operating Characteristic (ROC) curve

6. A.Sudha et al. [6] aims to implement algorithms like Decision trees, naïve Bayes classifiers, and neural networks to predict the possibility of a stroke. Thepatient dataset is collected from healthcare institutes that have symptoms of stroke disease. The stroke disease dataset is pre-processed to make it suitable for the mining process. The Pre-processing technique removes duplicate records, missing data, and noisy and inconsistent data. Principle Component Analysis is used. It deals with a huge amount of dataset and reduces it to a lower dimension. Following this, feature subset selection is used for feature reduction. It removes the irrelevant data and selects the data which are relatedto stroke disease. Classification algorithms are applied for training and testing data sets and their results are evaluated to determine the most significant geneset. Classification algorithms were used and then sensitivity and accuracy were determined

7. Cemil Colak et al. [7] predicted the outcome of stroke using knowledge discovery process (KDP) methods, Artificial Neural Networks (ANN), and Support Vector Machine (SVM) models. The records of 297 (130 sick and 167 healthy) individuals were acquired from the databases of the department of emergencymedicine. After feature selection, a multi-layer perceptron was used for the prediction Accuracy and AUC was determined. The findings of the current study pointed out that ANN had more predictiveperformance when compared with SVM in predicting stroke. The advantagehere is, it gives high accuracy and disadvantage is it is difficult to implement

8. Ahmet KadirArslan et al. [8] assesses different data mining approaches to theprediction of stroke. Dataset collected from Turgut Ozal Medical Centre, Inonu University, Malatya, Turkey, comprised the medical records of 80 patients and 112 healthy individuals with 17 predictors and a target variable. Support vector machine (SVM), stochastic gradient boosting (SGB), and penalized logistic regression (PLR) were the different algorithms employed. After implementing the various algorithms, model performance evaluation metrics used were accuracy, an area under the ROC curve (AUC), sensitivity, specificity, positive predictive value, and negative predictive value. The grid search method was used for optimizing the tuning parameters of the models. We can conclude that SVM produced the best predictive performance compared to the other models according to the majority of evaluation metrics. The advantage is Very high

accuracy obtained and multiple performance metrics were computed. The disadvantage of this paper is insufficient was information provided

9. Ohoud Almadani et al. [9] explains the process of predicting stroke using data mining. Strokedata set is obtained from Ministry of National Guards Health Affairs hospitals,Kingdom of Saudi Arabia. The data set consisted of 1004 attributes and for dimensionality reduction, PCA was used. Data splitting is done for the trainingand test set. A data mining model was built with 95%accuracy. JRip and MLP were used for creating the model. JRip had the highest accuracy.The advantage is efficient algorithms and the disadvantage is performance analysis was insufficient

10. Sheng-FengSung et al. [10] proposes a method for developing a stroke severity index (SSI) by using administrative data. Stroke severity was measured usingthe National Institutes of Health Stroke Scale (NIHSS). Two data mining methods and conventional multiple linear regression (MLR) were used to develop prediction models. The model had an 80% accuracy and efficiently predicted stroke.

Methodology

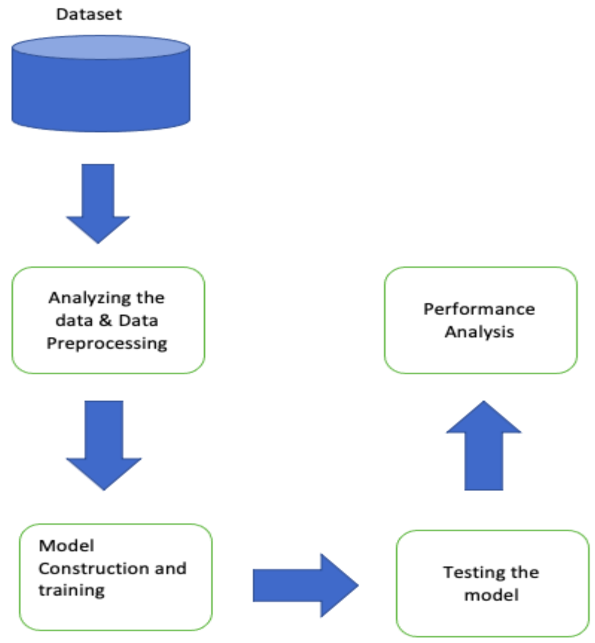

Dataset
Analyzing the data & Data Preprocessing
Performance Analysis
Model Construction and training
Testing the model

Figure 1. Methods flow

A. *Importing the data and visualising it*

Stroke prediction data set was collected from Kaggle for the implementation. The code was written in Python on a Jupyter Notebook interface. Some of the libraries used in this project are Pandas, Sklearn, Matplotlib, Seaborn and IPython. After importing the dataset, we compare the relations between the different attributes using Pair Plots from the Sea Born library. Some more data visualizations are carried out which helps us in analysing the dataset. This is carried out on RapidMiner.

B. *Data Pre-processing*

Following this, the first step towards creating the model is the Data Pre-processing step. Data pre-processing in this case involved checking for duplicate values, checking for null values and replacing them with the average value, converting string values to numeric values, removing unwanted classes, one hot encoding step and sorting out data imbalance issues.

C. *Construction of the model*

After the data is pre-processed, the model is built. Decision tree model is selected to be built as it is an efficient supervised learning algorithm for classification and prediction. The useful features are selected and then the data set is split into training and test data sets using appropriate functions where we distribute 80% of the data to the training set. After this, the data is scaled.

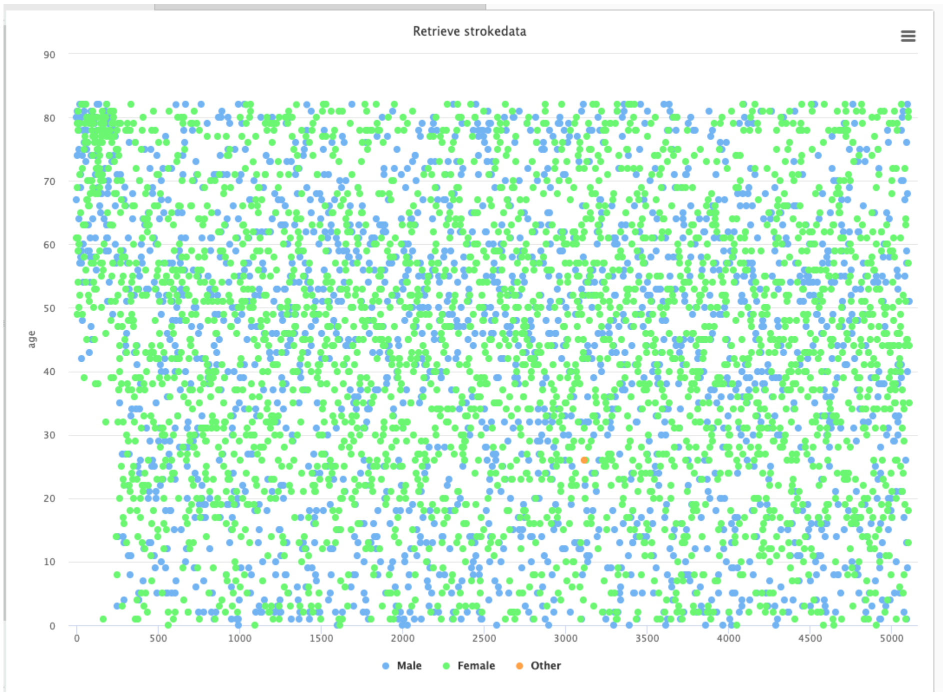


Figure 2. Age and gender relationship of the dataset used using scatter/bubble plot Image

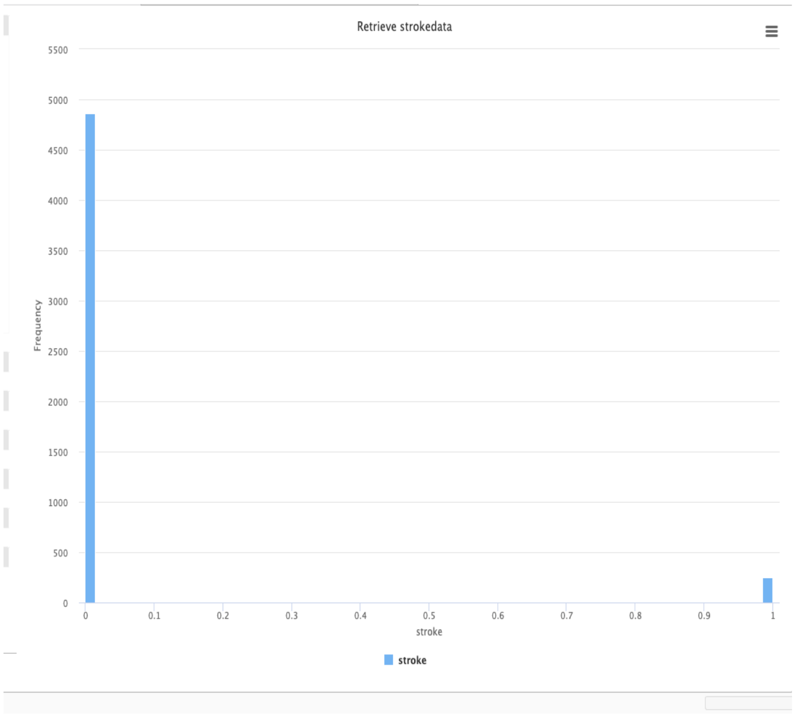


Figure 3. Stroke and non-stroke frequency count using histogram plot

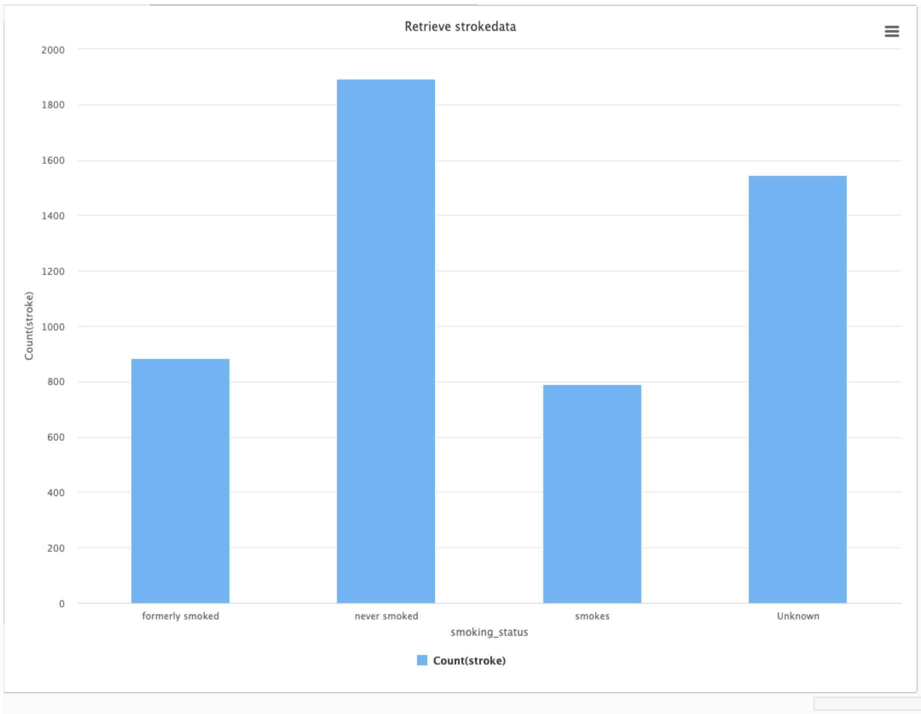


Figure 4. Relationship between smoking status and occurrence of stroke using Bar plot

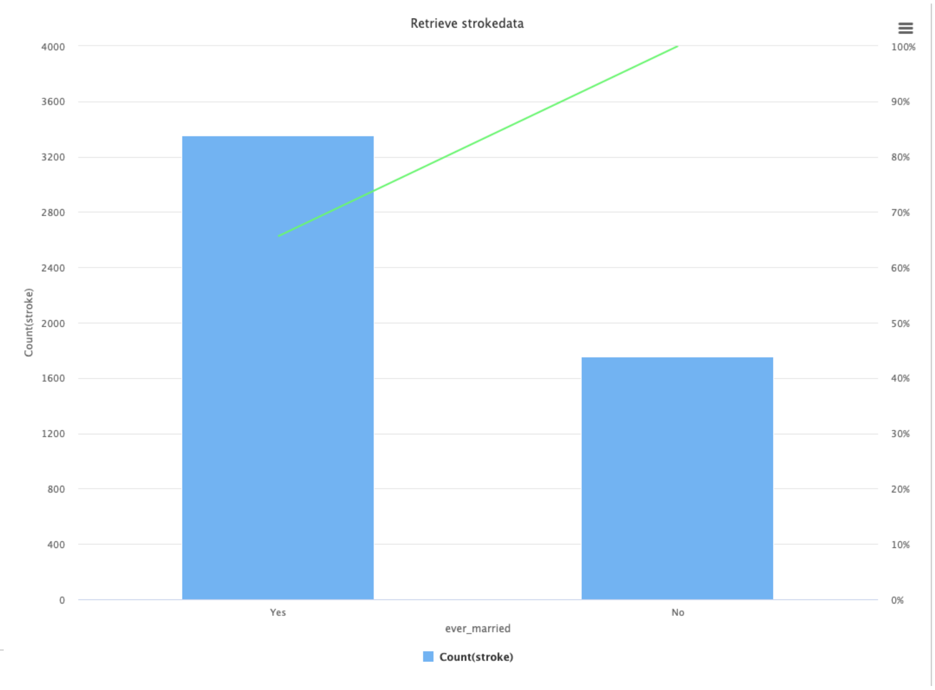


Figure 5. Relationship between marital status and stroke occurrence using Pareto

A correlation matrix is then built for further data visualizations.

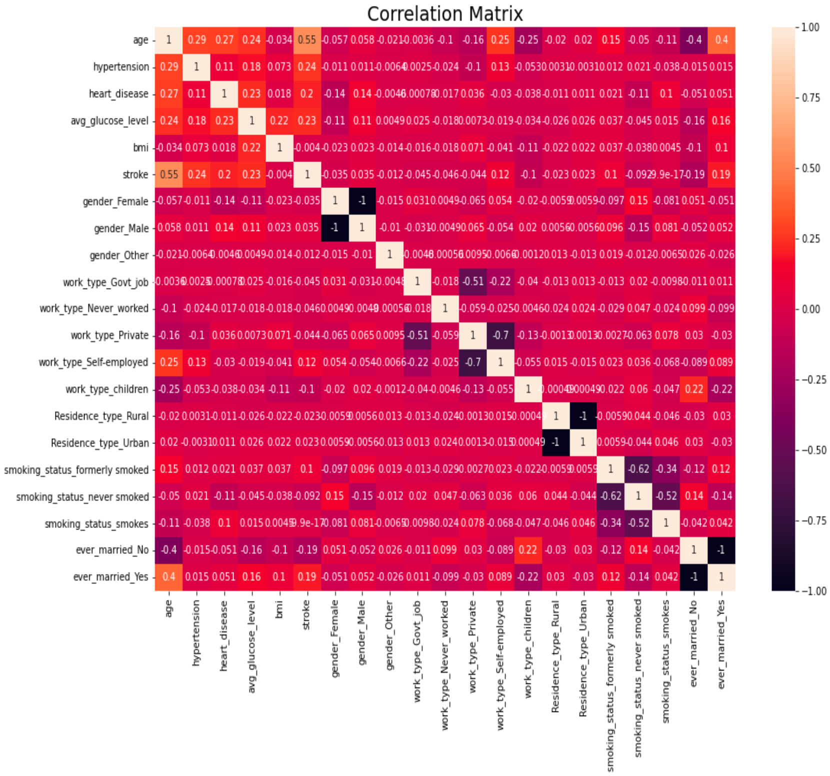
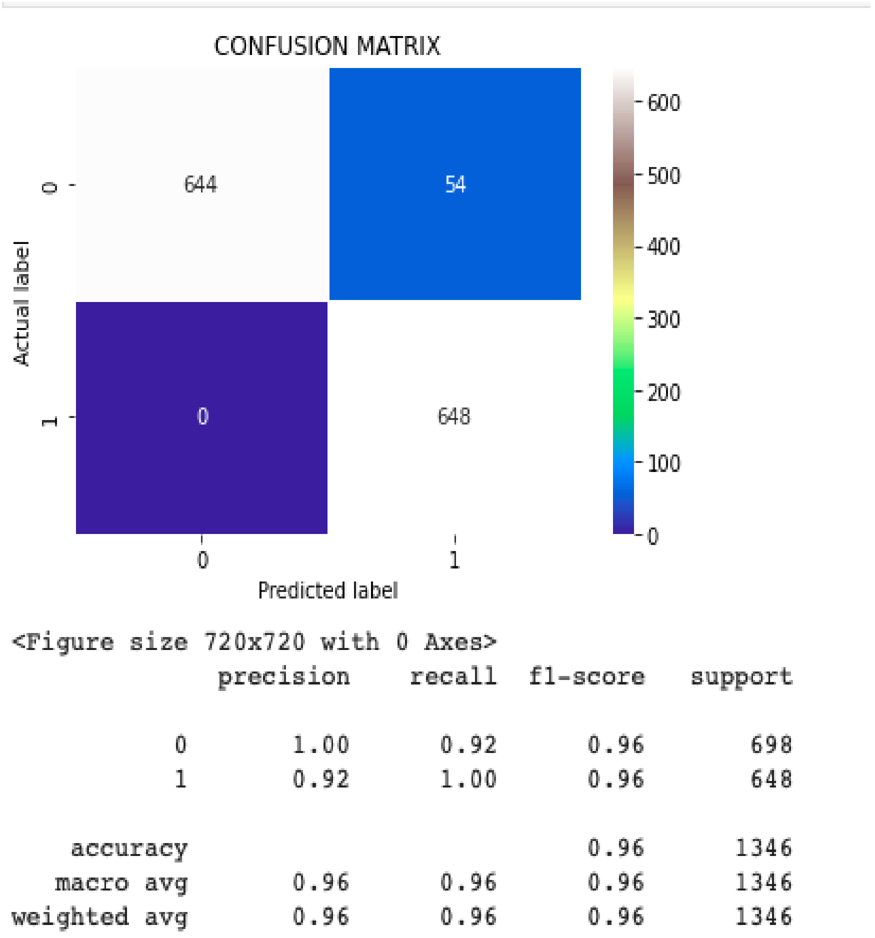


Figure 6. Correlation matrix of the pre-processed data set

The next step is creation of a decision tree. A decision tree is built with max depth 18 and we fit the training data. Accuracy and score are then calculated.

D. Performance Analysis of the model

The model produces a training score of approximately 0.97 and test score of approximately 0.95. The last step is the performance analysis by construction of confusion matrix and performance report.



```
<Figure size 720x720 with 0 Axes>
              precision    recall  f1-score   support

           0       1.00      0.92      0.96       698
           1       0.92      1.00      0.96       648

    accuracy                           0.96      1346
   macro avg       0.96      0.96      0.96      1346
weighted avg       0.96      0.96      0.96      1346
```

Figure 7. Confusion matrix and performance report

Results and discussion

The outcome of this paper is an efficient stroke prediction model which can predict stroke in a percent as accurately as 95%, taking in the details of the patients such as BMI, glucose levels and age. Although there have been previous researches on the same topic, this study proves to be comparatively simplified yet efficient. In contrast to multiple other studies making use of algorithms like KMeans and SVM models, Decision tree is the best fit for this challenge. In addition, data visualisation techniques used here help in gaining insight on how different lifestyle habits can also affect the chance of a person get attacked by a stroke. High accuracy available in this model makes it a reliable mean to dodge the possibility of a stroke attack. Absence of overfitting and underfitting and balanced training data makes it an efficient algorithm to varied kinds of patient data. Although this model provides a high accuracy to predicting stroke, it doesn’t provide information to the patient about the type of stroke he/she is susceptible to face and doesn’t enlighten the patient about which factor is causing the most risk to him/her and this has to be kept in mind when making use of the algorithm. This is a potential zone for future research and studies.

Conclusion

Stroke is an extremely harmful disease which causes millions of deaths, damages and disabilities all around the world. This problem can be tackled by efficient recognition and preventive measures. Machine learning models which predict stroke possibility are of great importance as it can even save lives. The motivation behind this report is to develop a model based on decision trees which can provide a patient the probability of him/her experiencing a stroke in the future. Decision trees are selected as it is a suitable algorithm for classification and predictive purposes. Upon construction, a model of training accuracy score of 0.97 and test accuracy score of 0.95 is achieved. Further, performance analysis is carried out by building a confusion matrix, providing numbers of precision, recall, f1 score and so on. A classification report is also computed. Several data visualisations are also carried out at different stages of the process.

References

[1] Amini,Leila,etal."Predictionandcontrolofstrokebydata

mining." International journal of preventive medicine 4.Suppl 2 (2013): S245.

[2]
M. S. Singh and P. Choudhary, "Stroke prediction using artificial
intelligence," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), 2017, pp.158-161, doi: 10.1109/IEMECON.2017.8079581.

[3]
R. S. Jeena and S. Kumar, "Stroke prediction using SVM," 2016 International Conference on Control, Instrumentation, Communicationand Computational Technologies (ICCICCT), 2016, pp. 600-602, doi: 10.1109/ICCICCT.2016.7988020.

[4]
M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, "Performance Analysis of Machine LearningApproaches in Stroke Prediction," 2020 4th International Conferenceon Electronics,
Communication and Aerospace Technology (ICECA), 2020, pp.
1464-1469, doi: 10.1109/ICECA49313.2020.9297525.

[5]
Sailasya, Gangavarapu, and Gorli L. Aruna Kumari. "Analyzing theperformance of stroke prediction using ML classification
algorithms." International Journal of Advanced Computer Science andApplications 12.6 (2021).

[6]
Sudha, A., P. Gayathri, and N. Jaisankar. "Effective analysis andpredictive model of stroke disease using classification
methods." International Journal of Computer Applications 43.14(2012): 26-31.

[7]
Colak, Cemil, Esra Karaman, and M. Gokhan Turtay. "Applicationof knowledge discovery process on the prediction of
stroke." Computer methods and programs in biomedicine 119.3(2015): 181-185.

[8]
Arslan, Ahmet Kadir, Cemil Colak, and Mehmet Ediz Sarihan."Different medical data mining approaches based prediction of ischemic stroke." Computer methods and programs in biomedicine 130 (2016): 87-92.

[9]
Almadani, Ohoud, and Riyad Alshammari. "Prediction of strokeusing data mining classification techniques." International Journal ofAdvanced Computer Science and Applications 9.1 (2018).

[10]
Sung, Sheng-Feng, et al. "Developing a stroke severity indexbased on administrative data was feasible using data mining techniques." Journal of clinical epidemiology 68.11 (2015): 1292-1300.