An Effective Stroke Prediction System using Predictive Models

Soodamani Ashokan[1], Suriya G.S Narayanan[2], Mandresh S[3], Vidhyasagar Bs[4], Paavai Anand G[5]

[1,2,3]Under Graduate, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India
[4,5]Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India

---***---

Abstract - *This paper provides an effective method for detecting stroke. R language in RStudio is used to conduct data analysis and for the construction of a prediction application. Making a random user be able to test themselves for stroke is the primary objective of this project. The web application and the data analysis parts are made to work on the patients' characteristics data. The stroke training and test datasets are gathered and exploratory data analysis is performed. The most efficient and accurate variables required to predict stroke in an individual is obtained through Feature Selection and as per the variables gained, the features which influence the disease prognosis is obtained. Predictive modeling is performed on this processed data with various classification models such as Random Forest, Decision tree, Logistic Regression and Support Vector Machines. The web application is made to process user inputs and predict the occurrence of stroke using the most accurate model.*

***Key Words*: Stroke, patient characteristics, predictive modeling, prediction, web application.**

1. INTRODUCTION

Neurological disorders deal damage to the central and peripheral nervous system. Some of these diseases can be treated whereas others cannot be. People, mainly after the age 60 suffer from neurological problems such as Alzheimer's disease and Parkinson's disease[1, 2], making age a significant factor for developing this disease. The causes differ from being genetic disorders, infections, lifestyle to any health problems that may affect the brain. There are more than 600 diseases of the nervous system, such as stroke, brain tumors, epilepsy and many more. Around 15 million people[3] suffer from stroke each year. There exits not much efficient means for the patient to predict whether he or she could possess such diseases and this research mainly focuses on that. An ease of use disease prediction application that works on properly analysed data is very much needed for any user to test and realise one's medical condition. This can help in taking the necessary treatments from a hospital. The application also serves useful in making the users aware of the need for a proper lifestyle. Around 1,654,577 people were admitted due to neurological conditions in the year 2016/17[4]. The neurological conditions include Autism, Dementia, Epilepsy and so on[4]. Men are found to have a stroke at a younger age than women and stroke related deaths occur more in women[5]. The dataset[6] used for this research belongs to the records of individuals tested for stroke with variables such as gender, age, lifestyle attributes, Body Mass Index, demographic regions and so on.

The estimated lifetime risk of the U.S. population as per gender and age for Alzheimer's disease(AD) in 2019 is displayed in the following figures. More percentage of women[7] were found to have the risk of AD than men as given in chart-1. People with ages 75-84 years (i.e) 2.6 million people, were affected with AD as shown in chart-2.

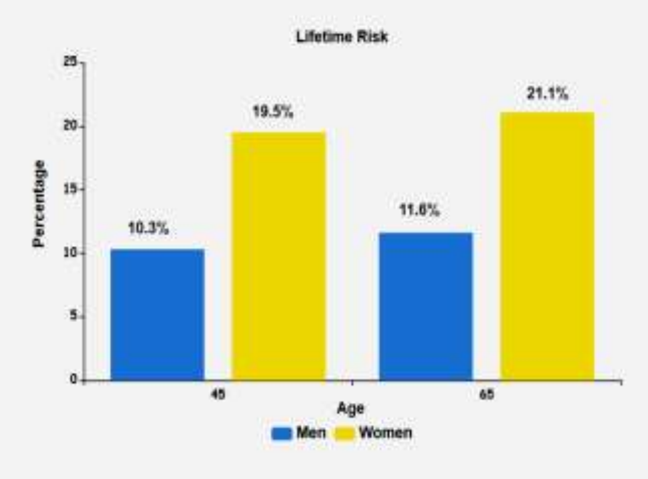


Chart -1: Alzheimer's lifetime risk

R Studio, a GUI for R programming language is used for working with the datasets and providing the required analysis. R and its packages are very helpful for conducting the analysis and presentations for understanding the analysis[8]. It helps in understanding and analysing the data in a statistical manner and performing operations. The first method for conducting the research involves cleaning and preprocessing of datasets to eliminate Not Applicable(NA) and empty values. A separate test dataset is used to test the accuracy of the predictive models and hence the test dataset is cleaned and preprocessed the same way as the training dataset.

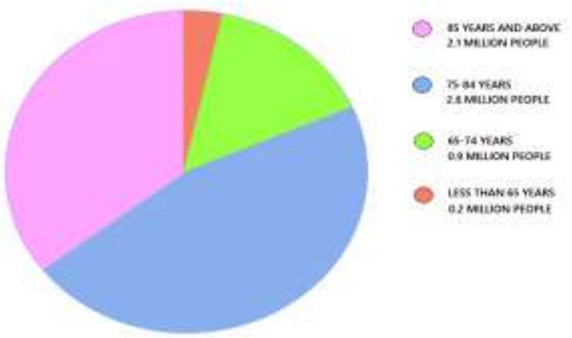


Chart -2: Ages of people with AD

Then exploratory data analysis, which includes plotting of the variables related to the particular disease, that is Stroke and testing the independence of the variables with respect to the disease using Chi-Square test[9]. The Feature Selection phase provides the necessary variables that we need for construction of our prediction system. Then latter methods include finding the accuracy of classification models by conducting predictive modelling. The predictive models which we build are Random Forest, Decision Tree and Support Vector Machine models[10]. Finally the prediction accuracy, sensitivity, specificity and computational time of the models used are found out. From this result, the best model out of the three is chosen for using it in our prediction application. The accurate model gained then shifts this research from data analysis to a web application construction[11]. The web app is mainly built using Shiny, which is a package of R[12]. This package provides scripts for writing the user interface(UI) part as well as writing a server part for the application and helps in prediction.

2. RELATED WORK

Researchers have worked on early diagnosis of neurological disorders, which can be hard to achieve. Their contributions and related work are presented here to understand the basis of the project. Ane Alberdi et al.(2018) had collected home behavior data of patients in order to detect the symptoms for Alzheimer's disease. The symptoms that they were able to obtain as an end result were related to the mood, cognition and mobility of patients. The researchers worked on a smart home solution[13] by which sensors can monitor the patients and help in detecting the multiple symptoms. The data assessed was from 29 older adults, who lived in smart homes and monitored for a duration of less than 1 month to 60 months. Regression models and classification models were also used on the data. This research helped in obtaining necessary behavioral features for detecting AD in patients and the changes in patients that could indicate the AD symptoms. Stroke is yet another neurological disease whose risk detection is a challenge. Yonglai Zhang et al.(2018) conducted a research on stroke patients to detect the risk of stroke[14]. The dataset worked upon consisted of medical tests and other archived data on 792 patients at a Beijing hospital. The data analysis was conducted in three stages- filter stage, voting stage and wrapper stage. The filtering was done based on the STD variable (standard deviation) and SVM algorithm was used for classification accuracy of feature subsets. The combined use of SVM with glow-worm swarm optimization algorithm was done to calculate the required features or conduct feature selection on the data. These features helped determine the major risk factors for stroke in patients.

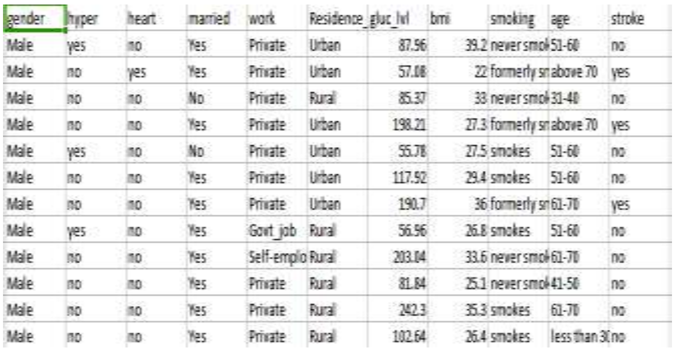
Farrikh Alzami et al.(2018) provided a feature selection method to classify seizures caused by epilepsy. The EEG dataset used for this research was provided by the University of Bonn, Germany. Hybrid feature selection was done by which subsets were obtained. The mRMR, Fisher, Chi-Square and Relief -F were applied for feature selection in this research. The subsets obtained were combined using rank aggregation. Furthermore the subsets obtained from aggregation were passed to a base classifier for obtaining the learning model and prediction. Finally voting was done for predicting the classification and detection tasks. The research[15] or AHFSE algorithm provided high performance when compared with other methods. Sudden unexpected death in epilepsy (SUDEP), is the death related to epilepsy. The researchers, Wanchat Threeanaew et al.,(2018) intended on discovering reliable physiological biomarkers from multimodal data[16]. Multimodal data had been collected from five different centers for more than 3 years. Feature extraction, feature reduction and other integrative analysis algorithms were used for analysis of the data.

Another research[17] of Alzheimer's disease includes the study by Pholpat Durongbhan et al.,(2018) which aimed to obtain biomarkers using Quantitative Analysis of Electroencephalography through a framework consisting of data augmentation, feature extraction, K Nearest Neighbour (KNN) classification, quantitative evaluation and topographic visualisation. Twenty HC and 20 AD subjects participated in this research and data was collected from them. MATLAB had been used for research purposes. The proposed framework was able to accurately classify the records and found important features as biomarkers for proper diagnosis of disease progression. A cure may not be possible for Alzheimer's disease (AD) but this paper provided another alternative, that is, the major risk factors leading to this disease was analysed and these factors helped prevent the disease. The data for this research was from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The dataset used consisted of subjects who undergo cognitive impairment and Alzheimer's Disease. The classification and ranking of the risk factors was done using machine learning models and classification techniques. The data used was from ADNI as well as some collected data on the risk factors for this research. Mohamed Mahyoub et al.(2018) proposed a framework for early detection of Alzheimer's, which had not been possible through existing research[18]. The framework proposed constructed a baseline dataset, deployed feature selection on it and increased the dataset variables as per the required and tested risk factors. The techniques were deployed in MATLAB and R studio. Deep Learning techniques were also used for modeling and formulation. A tool for early diagnosis of AD had been resulted from this method and could be used on various subjects and participants for checking and deciding if they had AD. This research solved the fact that AD can also be diagnosed at an earlier stage rather than later stages.

Our study provides insight on incorporation of data analysis into a prediction system and the significance in doing so. Making the analysis results to reach the user through an application is a different path taken up by this study rather than just stopping at data analysis.

3. DATASET DESCRIPTION

The data collection and analysis based on patient characteristics can be useful for our system user since any normal user can easily provide his/her general data. Complicated clinical data information such as Magnetic Resource Image scans and images are not supported by this prediction system and is created for the user to test oneself on the go. The dataset for stroke is collected from Kaggle, a dataset repository for data scientists and analysts. The dataset as given in figure-1 offers 12 variables consisting of patient id, gender, hypertension, age, heart disease rate, marriage status, occupation, type of residence, glucose level, patient BMI, smoking status and stroke yes/no prognosis.



gender	hyper	heart	married	work	Residence_	gluc_lvl	bmi	smoking	age	stroke
Male	yes	no	Yes	Private	Urban	87.96	39.2	never smok	51-60	no
Male	no	yes	Yes	Private	Urban	57.08	22	formerly sn	above 70	yes
Male	no	no	No	Private	Rural	85.37	33	never smok	31-40	no
Male	no	no	Yes	Private	Urban	198.21	27.3	formerly sn	above 70	yes
Male	yes	no	No	Private	Urban	55.78	27.5	smokes	51-60	no
Male	no	no	Yes	Private	Urban	117.52	29.4	smokes	51-60	no
Male	no	no	Yes	Private	Urban	190.7	36	formerly sn	61-70	yes
Male	yes	no	Yes	Govt_job	Rural	56.96	26.8	smokes	51-60	no
Male	no	no	Yes	Self-emplo	Rural	203.04	33.6	never smok	61-70	no
Male	no	no	Yes	Private	Rural	81.84	25.1	never smok	41-50	no
Male	no	no	Yes	Private	Rural	242.3	35.3	smokes	61-70	no
Male	no	no	Yes	Private	Rural	102.64	26.4	smokes	less than 3(	no

Fig-1: Stroke dataset

4. SYSTEM DESIGN

The system design is explained using figure-2. The patient dataset used for the system is collected from the stroke dataset source. Then the data undergoes a preprocessing phase, where unwanted and missing data is removed. This phase is also an important part because proper data analysis and filtering of data depends on preprocessing. Then the next part consists of exploratory data analysis where the variables are plotted and correlation of variables with disease prognosis is checked. This exploratory analysis is done for clear representation of data in a statistical format. The independent test (chi square test) is done to find the independency and dependency between the predictor and response variable. This research uses Chi Square test as a feature selection process too. Recursive Feature Elimination is another method followed for eliminating weak features from the data.

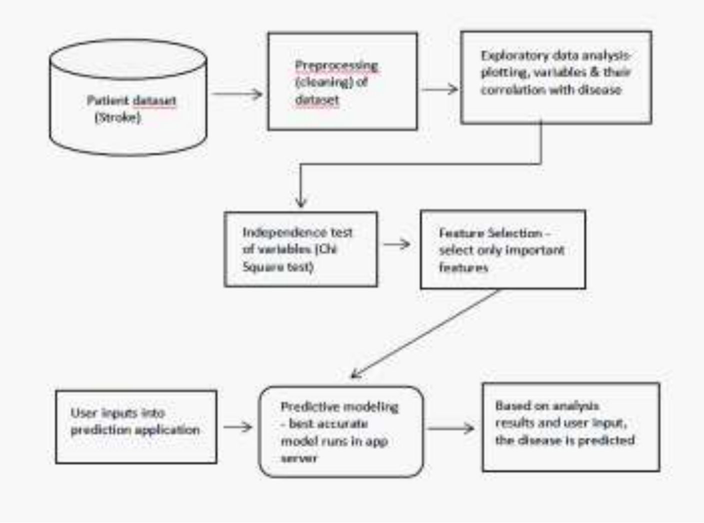

Patient dataset (Stroke)
Preprocessing (cleaning) of dataset
Exploratory data analysis- plotting, variables & their correlation with disease
Independence test of variables (Chi Square test)
Feature Selection - select only important features
User inputs into prediction application
Predictive modeling - best accurate model runs in app server
Based on analysis results and user input, the disease is predicted

Fig-2: Stroke prediction system design

Then the predictive modelling phase is done to select an accurate model for prediction of diseases. It involves building classification models using Logistic Regression, Decision Tree, Random Forest and Support Vector Machines algorithms and analysing the prediction accuracy of the models. This phase is the most significant part for the web application because the application server runs on the best model out of the four. Any user provides data into the prediction application, which gets converted into test data. Then the built model helps in predicting stroke on the user test data.

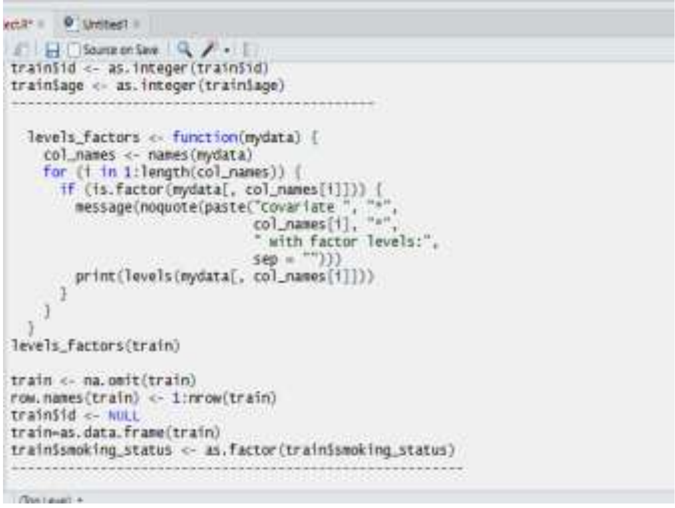
5. METHODOLOGY

The data analysis is carried out on the dataset in different phases. Various functions and libraries in R are used for this purpose. The respective phases and methods followed for this research are listed and explained in detail.

5.1 Preprocessing

Preprocessing is the initial step in conducting the analysis. This module provides elimination of unwanted values as well as cleaning the dataset. A dataset can be loaded with Not Applicable (NA) values or even empty values. These values need to be removed by removing the whole row which contains such a value. Zero values also can be neglected and the dataset needs to be rid of these values. This is where preprocessing comes in. This step is important mainly because the rest of the modules need to work upon cleaned data for proper analysis. The datasets are read into RStudio application as .csv file and the variables are processed into numerical and categorical variables for our study. After this step, the cleaning of variables is done. On cleaning the dataset with 43K observations has reduced to

23K observations given as in figure-3. Out of this cleaned data, only 6000 observations are taken into consideration for further analysis since lesser data can help in achieving a higher computational and accuracy rate. The result of this module is a cleaned dataset which is ready for exploratory analysis on it.



```
train$id <- as.integer(train$id)
train$age <- as.integer(train$age)
-------------------------------------------

  levels_factors <- function(mydata) {
    col_names <- names(mydata)
    for (i in 1:length(col_names)) {
      if (is.factor(mydata[, col_names[i]])) {
        message(noquote(paste("Covariate ", "*",
                              col_names[i], "*",
                              " with factor levels:",
                              sep = "")))
        print(levels(mydata[, col_names[i]]))
      }
    }
  }
levels_factors(train)

train <- na.omit(train)
row.names(train) <- 1:nrow(train)
train$id <- NULL
train=as.data.frame(train)
train$smoking_status <- as.factor(train$smoking_status)
------------------------------------------------------
```

Fig-3: Preprocessing of datasets

5.2 Exploratory Data Analysis

When cleaned data is obtained, data exploration must be done in order to get insight from it. Analysis of the variables is done in this phase in order to get valuable information about the explanatory variables and their relation with the response variable. Yes/no prognosis of stroke and depression act as the response variable in both datasets for this process. The aim of this phase is to display and get information on how the other variables are influencing and relating with our response variable through plots and visualisation. The different variables of stroke dataset such as age, BMI, glucose levels, gender and so on are checked for their relation with the stroke prognosis variable. On analysing age through box plot, patients with ages 60 to 77 are found to have stroke than the patients at other ages and the mean age for stroke is found to be 67. Patients with a BMI of 26 upto almost 35 are found to get affected with stroke. This shows that overweight and obese people are most likely to get affected by stroke than other BMI levels. Information like this can only be achieved through the EDA phase and hence its importance.

The average glucose levels for a patient with stroke is seen to be between 95 to 120. The above three variables are numerical variables and that is why a box plot and histogram are used for analysing them. Rest of the variables that we are going to work upon are categorical variables i.e factors. Barplots are used for plotting these variables efficiently. Male patients are found to possess stroke than the female ones. 72% of the patients with heart disease appear to have strokes more than people who do not have heart disease. Apart from these variables, other variables like smoking status and high blood pressure or hypertension levels seem to be influencing stroke. Subjects who formerly smoked have a higher chance of getting stroke followed by the current smokers and non-smokers. Increase in blood pressure also increases the chances of stroke as given in figure-4. Self-employed people, people staying in urban areas and married people are found to possess more strokes than people with government jobs, private jobs, people staying in rural parts and single people.

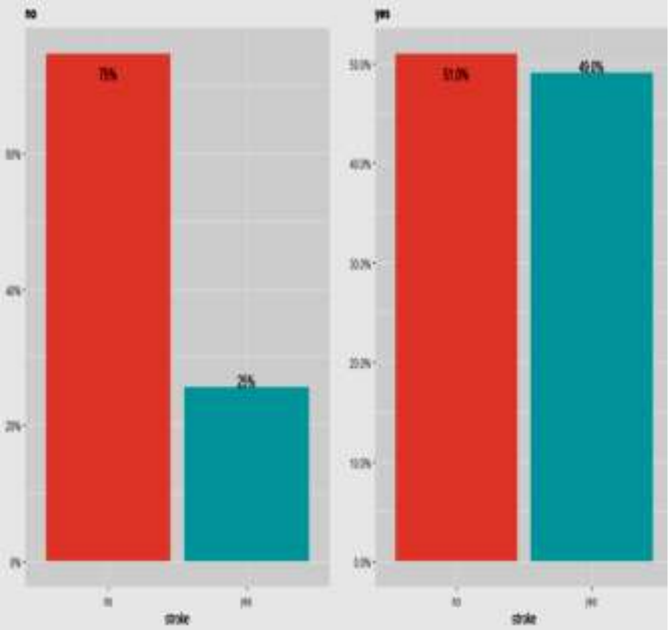


Fig-4: Bar plot of hypertension grouped by stroke

5.3 Feature Selection

Feature selection is the process in which we select the most important features out of all the independent variables. The features are selected for our study based on two feature selection methods[19]- Filter method and Wrapper method. The Filter method that we use is Chi-Square test and the Wrapper method includes Recursive Feature Elimination. Chi-Square method tests the independence of the variables whereas the latter method provides a process for eliminating the less important features until only the best features are chosen from the dataset.

Chi-Square test tests whether the independent variables or predictors satisfy the null hypothesis or the alternative hypothesis. Null hypothesis states that the predictors and the outcome variable are independent whereas in alternative hypothesis, they are dependent. If the alternative hypothesis is satisfied, then the feature or variable is selected else feature is neglected. Recursive Feature Elimination helps in selecting only the best and optimal features through cross validation and continuously eliminating weak features. Variable importance is also checked on the dataset using varImp() function in R. The methods showed that the

Residence variable to be the least important variable out of the ten variables in the dataset.

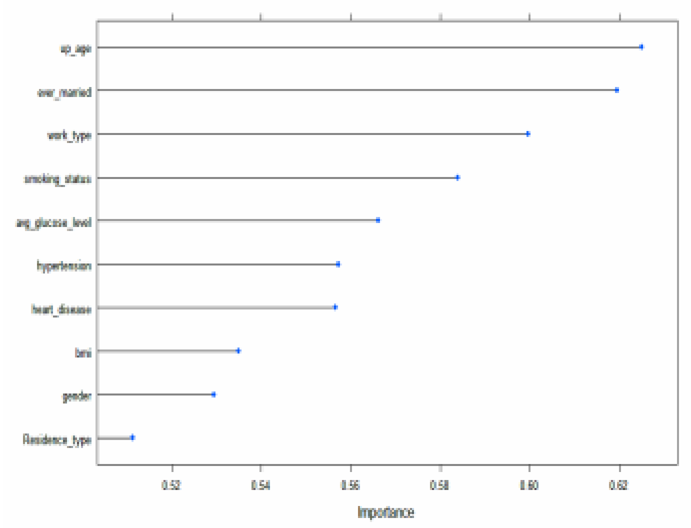


Fig-5: Variable importance plot

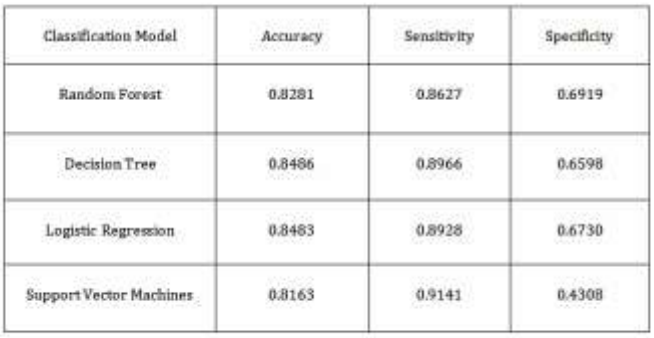
5.4 Predictive Modeling

This phase follows the feature selection methods by which the best nine features - gender, hypertension, age, heart disease rate, marriage status, type of work, glucose level, patient BMI, smoking status are selected. The training and the test dataset consists of nine features each. The test dataset does not contain the outcome variable i.e. stroke variable and the main aim of this phase is to predict this variable by fitting classification models into the training dataset. The classification models used for this study include Random Forest, Decision Tree, Logistic Regression and Support Vector Machines. The accuracy of the models are checked individually and the most accurate model is chosen for constructing the stroke prediction web application.

Some of the R libraries needed for classification are caret. randomForest, e1071, rpart and rpart. plot. Rpart function is used to create a decision tree classifier for the training dataset. The rpart. plot function helps in plotting a decision tree based on training data. From the tree, it is noted that people below the age 60 do not have strokes. The people with ages greater than 60 and with a smoking habit are found to have been affected by stroke. The svm, randomForest, glm functions are used for creating the SVM classifier, Random Forest classifier and Logistic Regression classifier respectively.

The SVM is built as a classification type machine along with a polynomial kernel for predicting. A confusion matrix is constructed for all the models in order to evaluate the performance of the models based on accuracy. Decision tree provided the highest accuracy followed by Logistic Regression, Random Forest and Support Vector Machines as shown in the following table. Now the prediction application server is built using the Decision Tree model and the prediction is made.

Table -1: Accuracy differentiation among models

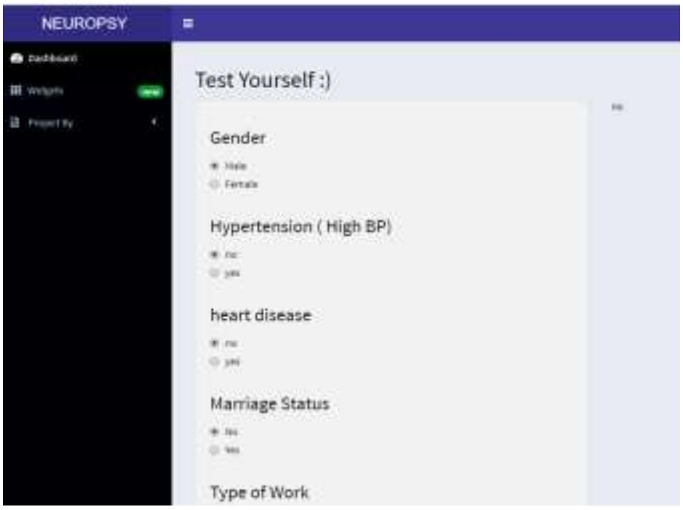


Classification Model	Accuracy	Sensitivity	Specificity
Random Forest	0.8281	0.8627	0.6919
Decision Tree	0.8486	0.8966	0.6598
Logistic Regression	0.8483	0.8928	0.6730
Support Vector Machines	0.8163	0.9141	0.4308

6. APPLICATION

Shiny package of R is used to create web applications which can run in localhost or can be deployed in the web itself. Shiny provides a User Interface and Server as separate files or both UI and server in a single file called app.R. The user interface code for the prediction application consists of a Shiny dashboard, which provides an application-like appearance to our app and is well presented. The rest of the UI code consists of displaying radio buttons and numeric input boxes so that the user can select from input options which correspond to the stroke dataset variables.

The user interface helps in getting the inputs from the user and making test data out of it. The best model which was chosen in predictive modeling comes into play here. Since Decision Tree showed more accuracy than the other classification models, it is used in the app server. The model is fit into the stroke training dataset and the classifier is built. It is now tested with the test data obtained from the user. This is the reason for choosing an accurate model to build the prediction application as the classifier can effectively predict the outcome variable for the user data. This outcome variable predicted is displayed to the user as a stroke yes/no. As the user gives his/her inputs, the application is able to dynamically give a prediction in figure-6.



Test Yourself :)

Gender

● Male
○ Female

Hypertension (High BP)

● no
○ yes

heart disease

● no
○ yes

Marriage Status

● No
○ Yes

Type of Work

Fig-6: Shiny prediction application

7. RESULTS AND ANALYSIS

The training accuracy for the four models are found out by partitioning the training dataset itself into training and validation sets. The training accuracy produced by the classification models are found to be only eight to six percent higher than our test accuracy given in table-1. This gives us insight on the better performance of the trained models. When running the application and checking for results, the application is seen to provide the exact accurate results given by the Decision Tree model. Users having ages above 60 are seen to have higher chances of getting stroke. Users with age 61-70 do not have stroke if they have never smoked and those above 70 have chances of stroke, no matter the inputs given. This prediction shows that after an age of 60 and above, stroke is most likely to affect people. Thus the prediction is in accordance with well-known facts about stroke.

Even though labeled data is used for the prediction, being able to analyse unlabeled stroke related data can provide deep knowledge on cases of stroke. Association Rule Mining is done in our research to find out unknown rules and their subsets, which cannot be known through supervised learning. This analysis can help in further research dealing with unlabeled data. Arules, arulesViz and apriori function is used for this rule mining process. The rules generated are fine tuned using high level confidence and support parameters. By doing so, it is found that married people above 70 years of age or people above 70 without hypertension and heart disease also have the possibilities of getting stroke.

8. CONCLUSIONS

The incorporation of data analysis with a web application is successfully done with this study. The application shows that the data analysis results can also be made to reach the users and help them out with stroke prediction. This type of integration helps the user to easily understand the plots and statistical work taking place in data analysis. Further developments in this research can include predicting various other diseases apart from stroke. The complexity of the system can also be increased by taking medical scans and images as user inputs and trying to predict the disease. The application can be made more user friendly too by recording user inputs and information in a database and providing valuable suggestions to users as per their disease.

REFERENCES

[1] "What is Alzheimer's Disease? Symptoms & Causes |alz.org," Alzheimer's Association.[Online]. Available:http://www.alz.org/alzheimers-dementia/what-is-alzheimers.

[2] "Parkinson's disease-Symptoms and causes," Mayo Clinic,2018.[Online].Available:https://www.mayoclinic.org/diseases-conditions/parkinsons disease/symptoms-causes/syc-20376055.

[3] "Stroke Statistics," The Internet Stroke Center. [Online].Available:http://www.strokecenter.org/patients/about-stroke/stroke-statistics/.

[4] "Neuro Numbers 2019 - Neurological Alliance," The Neurological Alliance,2019. [Online]. Available:https://www.neural.org.uk/assets/pdfs/neuro-numbers-2019.pdf.

[5] "State of the Nation: stroke statistics | Stroke Association," Stroke Association,2018.[Online]. Available:https://www.stroke.org.uk/sites/default/files/state_of_the_nation_2018.pdf.

[6] "Healthcare Dataset Stroke Data," Saumya Agarwal, Kaggle,2018.[Online].Available:https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data.

[7] "2019 Alzheimer's Disease Facts and Figures Report," Alzheimer's Association,2019. [Online].Available:https://www.alz.org/media/documents/alzheimers-facts-and-figures-2019-r.pdf.

[8] M. Prakash, G. Padmapriy et al.,"A Review on Machine Learning Big Data using R," Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies, IEEE, 2018.

[9] Chris Chatfield,"Exploratory data analysis," European Journal of Operational Research,vol. 23, issue1,pp.5-13,1986.

[10] Bramesh S M, Puttaswamy B S et al., "Comparative Study of Machine Learning Algorithms on Census Income Data Set," Bramesh S M Journal of Engineering Research and Application, IJERA, vol. 9, issue 8,pp.78-81,2019.

[11] Shratik J. Mishra , Albar M.Vasi et al., "GDPS - General Disease Prediction System," International Research Journal of Engineering and Technology (IRJET), vol.05, issue 03, 2018.

[12] Shiny package-R Documentation.[Online]. Available:https://www.rdocumentation.org/packages/shiny/versions/1.4.0.2.

[13] Ane Alberdi, Alyssa Weakley et al., "Smart home-based prediction of multi-domain symptoms related to Alzheimer's Disease," IEEE Journal of Biomedical and Health Informatics, 2018.

[14] Yonglai Zhang, Wenai Song et al., "Risk Detection of Stroke Using a Feature Selection and Classification Method," IEEE Access, vol. 6, pp. 31899-31907, 2018.

[15] Farrikh Alzami, Juan tang et al., "Adaptive hybrid feature selection-based classifier ensemble for epileptic seizure classification," IEEE Access, vol. 6, pp. 29132-29145, 2018.

[16] Wanchat Threeanaew, James MacDonald et al., "Collection and Analysis of Multimodal Data for SUDEP Biomarker Discovery," IEEE Sensors Letters, vol.3, no.1, 2019.

[17] Pholpat Durongbhan, Yifan Zhao et al., "A Dementia Classification Framework using Frequency and Time-frequency Features based on EEG signals," IEEE Transactions on Neural Systems and Rehabilitation Engineering, pp. 1-10, 2018.

[18] Mohamed Mahyoub, Martin Randles et al., "Effective Use of Data Science Toward Early Prediction of Alzheimer's Disease," IEEE 20th International Conference on High Performance Computing and Communications,IEEE 16th International Conference on Smart City, IEEE 4th Intl. Conference on Data Science and Systems, pp. 1455-1461, 2018.

[19] "Intro to Feature Selection Methods for Data Science," towards data science,2019.[Online]. Available:https://towardsdatascience.com/intro-to-feature-selection-methods-for-data-science-4cae2178a00a.