

**School of Computer Science and Engineering**

**J Component report**

**Programme : BTech**

**Course Title : Foundations of Data Analytics**

**Course Code : CSE3505**

**Slot : F1**

**Title: BRAIN STROKE PREDICTION**

**Team Members:**

**Vaishnavi Sreedhar Jawalkar | 20BCE1571**

**Srividhya S | 20BPS1119**

**Yerramalli Sai Sreekar | 20BCE1296**

**Faculty: Dr. A Sheik Abdullah**   **Sign:**

**Date:**

**ABSTRACT**

The negative impact of stroke on society has led to a concerted effort to improve the treatment and diagnosis of stroke. With greater synergy between technology and medical diagnostics, caregivers are creating opportunities for better patient management by systematically extracting and archiving patients' medical records. Therefore, it is important to study the interdependence of these risk factors in medical records and to understand their relative contribution to stroke prediction. This project systematically analyses the different factors in electronic medical records for effective stroke prediction. Using various statistical techniques and principal component analysis, we identified the most important factors in stroke prediction.

We conclude that age, heart disease, average glucose level and hypertension are the most important factors in detecting stroke in hospitalized patients. Also, a random forest algorithm using these four attributes offers the highest accuracy rate and the lowest error rate compared to using all available input functions and other benchmarking algorithms. Since the dataset is very imbalanced in terms of stroke incidence, we present our results in a balanced dataset constructed by subsampling techniques.

Brain stroke by WHO is defined as the rapidly developed clinical signs of focal (or global) disturbance of cerebral function, lasting more than 24 hours or leading to death. Stroke is caused when the blood flow to the brain is blocked and causes deficiency of oxygen and nutrients and leads to death of brain cells in fraction of time. Brain strokes are also known as Brain attack. The purpose of this research is to create an accurate model for predicting Stroke based on the mentioned algorithms.

We have used R studio to implement to approach and solution for above objective. Methods are the algorithm which we have used to work on our dataset. We have used Generalized Linear Model -Logistic Regression -Dummy Var, Principal Components Analysis -Dummy Var, Generalized Linear Model - Logistic Regression -Dummy Var\_PCA, Data Split using Oversampling Data, Random Forest Model-Oversampling Data and Random Forest Model - Original Data

We explored and analysed how these different algorithms produce different results with varying accuracies. We realised the importance of data analytics and machine learning in the healthcare sector, we looked at how these algorithms might perform for other medical emergencies like heart attack and cancer. We saw the potential of R data analytic algorithm in determining and predicting medical emergencies and how if implemented in applications will help the community.

**INTRODUCTION**

A ‘mind stroke’ or a ‘mind attack’ is an occasion and circumstance that lots of us realize little about, however every one folks has the potential to save you and detect. When the mind is disadvantaged of blood and the oxygen it carries, or whilst bleeding inundates surrounding tissue and reasons the mind to swell, its effective operation will become compromised. Both incidents can purpose lasting vision problems, seizures, fatigue, lack of speech, reminiscence loss, and paralysis among different unfavourable effects. If excessive enough, they also can purpose death.

In low-and middle-profits countries, which encompass the ones of the WHO South-East Asia Region, over eleven million strokes arise each year. This reasons four million deaths annually, and leaves about 30% of survivors significantly disabled. For the 70% of survivors who recover, the probability of struggling similarly strokes is substantially increased. People with excessive blood stress, excessive cholesterol, coronary heart sickness, diabetes or a excessive blood-sugar stage are susceptible to mind stroke. So are human beings which can be obese, smoke or devour alcohol in massive volumes and are bodily inactive. In addition, the chance of stroke will increase with age, while men are much more likely to go through a stroke than females. Most of the lifestyle associated dangers may be decreased to save you mind stroke. People who smoke need to quit, and people who drink closely need to cease.

These elements by myself considerably multiply the probability of stroke. A eating regimen excessive in veggies and fruit and coffee in salt need to be consumed. Doing so will lower fatty deposits withinside the arteries which can purpose blockages, as properly as lessen the possibility of burst vessels that excessive blood stress brings. Regular workout needs to be undertaken - as a minimum half-hour of moderate intensity cardio pastime as a minimum 5 instances a week. Blood stress, blood-sugar and levels of cholesterol need to be checked regularly, with related conditions controlled in session with a fitness care issuer This easy however effective conduct can assist save you mind stroke and different noncommunicable sicknesses such as coronary heart sickness and diabetes. Identifying stroke’s early caution symptoms and symptoms is equally critical to save you incapacity or death.

Health structures must be in a function to behave decisively. The World Health Organization (WHO) estimates that fifteen million human beings international be afflicted by strokes every year, with one character death each 4 to 5 mins withinside the affected population. Stroke is the sixth leading cause of mortality in the United States according to the Centres for Disease Control and Prevention (CDC). Stroke is a noncommunicable disease that kills approximately 11% of the population. In the United States, about 795,000 human beings be afflicted by the disabling results of strokes on a normal basis. It is India’s fourth main purpose of death. Strokes are classified as ischemic or haemorrhagic. In a chemical stroke, clots obstruct the drainage; in a haemorrhagic stroke, a weak blood vessel bursts and bleeds into the brain. Stroke can be averted via way of means of main a healthful and balanced life-style that consists of abstaining from bad behaviours, which includes smoking and drinking, preserving a healthful frame mass index (BMI) and a median glucose level, and preserving an exceptional coronary heart and kidney function.Stroke prediction is critical and should be dealt with directly to keep away from irreversible harm or loss of life. With the improvement of era withinside the clinical sector, it's far now viable to expect the onset of a stroke via way of means of making use of ML techniques. The algorithms protected in ML are useful as they permit for correct prediction and right evaluation. The majority of preceding stroke-associated studies has targeted on, amongst different things, the prediction of coronary heart attacks. Brain stroke has been the problem of only a few studies. The most important motivation of this paper is to illustrate how ML can be used to forecast the onset of a mind stroke. The maximum vital factor of the techniques hired and the findings accomplished is that some of the 4 wonderful category algorithms tested, Random Forest fared the best, reaching a higher accuracy metric in contrast to the others. One disadvantage of the version is that it's far skilled on textual facts as opposed to actual time mind images. The implementation of ML category techniques the usage of R and facts evaluation has been done.

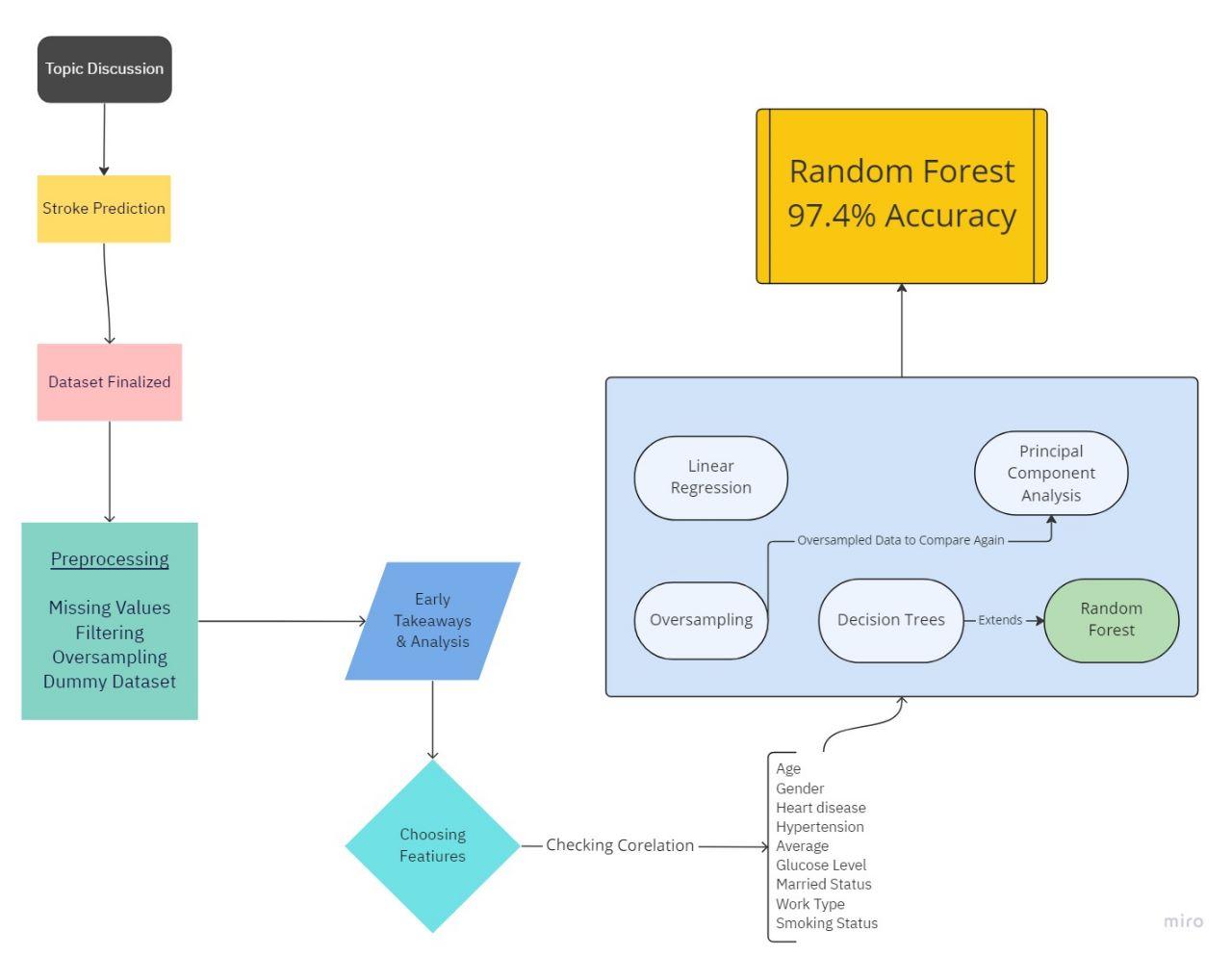
According to World Health Organization, stroke is the second main purpose of loss of life globally which approximates to 11% of general deaths. In the US, stroke is the main purpose of long-time period grownup incapacity and 5th main purpose of loss of life which approximates to 795,000 stroke occasion each 12 months with approximately 75% of them being the first-time strokes. Study suggests that strokes may be averted 80% of the time, via right education, on symptoms and symptoms of the stroke. Hence, it’s important that hospices actively take part in Stroke prediction to take away remedy delays and enhance results for brand new and acute stroke patients.

Stroke is a clinical circumstance which can cause the loss of life of a person. It’s a extreme circumstance and if dealt with on time we will shop one’s lifestyles and deal with them well. There can be n number of factors that can lead to strokes and in this project, we will try to analyse a few of them. We have taken the dataset from Kaggle. It has 11 variables and 5110 observations.

**LITERATURE REVIEW**

* (April 2022) Mrs. Neha Saxena, Mr. Arvind Choudhary, Mr. Deep Singh Bhamra, Mr. Preet Maru et.al [1] in their paper, “Brain Stroke Prediction using Machine Learning” used ML algorithms like Logistic regression, SVM, KNN, Decision Tress and Random Forest to determine and predict the risk of Brain Strokes. In this study, simple and efficient Graphical User Interface is designed to facilitate users and patients. This study obtained results which has a highest accuracy of 98.56%.
* (2021) Gangavarapu Sailasya, Gorli L Aruna Kumari et.al [2] presented the paper, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms” used machine learning algorithms like Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbours, Support Vector Machine and Naïve Bayes Classification. Naïve Bayes is the best with accuracy of 82%(approx.).
* (December 2020) Vamsi Bandi, Debnath Bhattacharyya, Divya Midhunchakkravarthy et.al [3] “Prediction of Brain Stroke Severity Using Machine Learning” Stroke Prediction algorithm is proposed by using the improvised random forest to analyse the risk factors of stroke. The Stroke Predictor model has an accuracy of 96.97% when compared with the existing models.
* (Oct 2021) Tahia Tazin, Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis and Mohammad Monirujjaman Khan et.al [4] in their paper, Machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, and Voting Classifier Technique were used in the study "Stroke Disease Detection and Prediction Using Robust Learning Approaches." The best model had a 96 percent accuracy rate and was Random Forest (approx.). The open-access Stroke Prediction dataset was used in this study.
* (October 2022) Nitish Biswas, Khandaker Mohammad Mohi Uddin, Sarreha Tasmin Rikta, Samrat Kumar Dey et.al [5] “A comparative analysis of machine learning classifiers for stroke prediction: Using 11 ML algorithms, including Support Vector Machine, K-Nearest Neighbors, Random Forest, Decision Tree, Naive Bayes, Logistic Regression, Adaptive Boosting, Gradient Boosting, Multi-Layer Perceptron, Nearest Centroid, and Voting Classifier, this paper took a predictive analytics approach. SVM and RF among the eleven classifiers exhibit the highest accuracy rates of 99.99% and 99.87%, respectively.
* (March 2022) Jin Qin, Lin Liu, Xu D. Su, Bin B. Wang, Bao S. Fu, Jun Z. Cui, Xiao Y. Liu et.al [6] in their paper, “The effect of PCSK9 inhibitors on brain stroke prevention: A systematic review and meta-analysis” used proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitors and tested the effect of them on Brain stroke using Statistical Analysis.
* (March 2022) Mandeep Kaur, Sachin R. Sakhare, Kirti Wanjale and Farzana Akter et.al [7] presented the paper, “Early Stroke Prediction Methods for Prevention of Strokes” used Deep learning and Neural Network algorithms like FFNN, LSTM, biLSTM and GRU for earlier prediction of brain stroke based on the bioelectric signals received from the brain. GRU shows high specificity and sensitivity values and hence is the best.
* (September 2022) Oscar Hoekstra, William Hurst, Joep Tummers in their paper [8] “Healthcare related event prediction from textual data with machine learning: A Systematic Literature Review” used neural network algorithms like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Conditional Random Fields (CRF) and NLP algorithms such as word2vec, BERT. CNN algorithm has the highest F1 score of 98.5% and it is considered to be more accurate.
* (May 2022) A.H. Alanazi, A. Cradock, L. Rainford et.al [9] in their paper, “Development of lumbar spine MRI referrals vetting models using machine learning and deep learning algorithms: Comparison models vs healthcare professionals” used Five algorithms were used: SVM, LR, RF, CNN, and Bi-LSTM. Both original data set and data set after pre-processing and cleaning is used. Comparatively, Bi-LSTM gave high accuracy.
* (November 2020) V. Jackins, S. Vimal, M. Kaliappan and Mi Young Lee [10] presented their paper on, “AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes” used ML algorithms like K-Means, Random Forest, Naive Bayes and DBSCAN algorithm for early prediction of diabetes, cancer and other heart diseases by finding correlation coefficient and using confusion matrix based on train and test accuracy. Accuracy outcome of Random Forest is greater and hence is the best for the analysis.
* (August 2020) Mecit Can Emre Simsekler, Abroon Qazic, Mohammad Amjad Alalamia, Samer Ellahhamd and Al Ozonoff [11] in their paper.Hospital management for improving patient safety using ML algorithms like the Random Forest Algorithm is discussed in "Evaluation of patient safety culture using a random forest algorithm." Using the examples of Portugal, France, Scotland, Palestine, etc., analysis was conducted.
* (May 2020) Muhammad Fazal Ijaz, Muhammad Attique and Youngdoo Son et.al [12] “Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods” used Chi-square feature selection method, Outlier detection method (DBSCAN, iForest) and Over-sampling method (SMOTE, SMOTETomek, Random Forest) to predict cancer using Biopsy results. Also developed a mobile app called “Cervical Cancer Prediction App (CCPM). DBSCAN + SMOTETomek + RF combination gave highest accuracy.
* (March 2021) Raneem Qaddoura, Ala’ M. Al-Zoubi, Iman Almomani and Hossam Faris et.al [13] “A Multi-Stage Classification Approach for IoT Intrusion Detection Based on Clustering with Oversampling” focussed on Intrusion detection of IoT-based data using K means algorithm and Over-sampling algorithm. Divided into clusters by K means clustering algorithm, the data has over-sampling ratio of 0.9 which was calculated using SMOTE.
* (November 2020) Wenjing Ye, Weiwei Lu, Yanping Tang, Guoxi Chen, Xiaopan Li, Chen Ji, Min Hou, Guangwang Zeng, Xing Lan, Yaling Wang, Xiaoqin Deng, Yuyang Cai, Hai Huang and Ling Yang, in their paper [14] “Identification of COVID-19 Clinical Phenotypes by Principal Component Analysis-Based Cluster Analysis” used Principal Component Analysis Algorithm to classify COVID 19 patients amongst the normal ones. Age is not the only factor to assess a patient’s condition.
* (2020) Carlo Ricciardi, Antonio Saverio Valente, Kyle Edmunds, Valeria Cantoni, Roberta Green, Antonella Fiorillo, Ilaria Picone, Stefania Santini and Mario Cesarelli et.al [15] Principal Component Analysis and Linear Discriminant Analysis Algorithm were used in "Linear Discriminant Analysis and Principal Component Analysis to Predict Coronary Artery Disease" to forecast CAD. The findings indicated that when LDA and PCA were combined, accuracy was high (98.4%).

**PROPOSED METHODOLOGY**



**EXPERIMENTAL RESULTS AND DISCUSSION**

When assessing risk factors for a stroke, the exploratory data analysis and data cleansing stages produced a number of insights and expectations for the future.

Early research suggested a connection between having strokes and having average blood glucose levels that are higher.

When compared to the data set mean of 106.15, the mean Average Glucose Level of stroke patients was 132.54. Heart disease and hypertension were also present in 18.88% and 26.51% of stroke patients, respectively, suggesting that these conditions are also likely risk factors for strokes.

General Linearized Modelling - Linear Regression - 94.7% Accuracy

The model discovered that factors such as age, hypertension, typical glucose levels, and whether or not a person worked for a private company all had statistically significant effects on the likelihood that they would get a stroke. The association between working for a private corporation and the risk of having a stroke is the new information in this situation. In actuality, the model discovered that working for a private corporation significantly raised the risk of stroke compared to other occupations.

According to exploratory data analysis, there is a clear link between ageing and strokes, as well as between having hypertension and high blood sugar levels and having a stroke. Despite the poor model accuracy, this makes it relatively simple to accept these findings as risk factors.

Principal Components Analysis: 94% Accuracy

The number of Principal Components to include in the model to obtain a specific degree of variation the model explains is shown by the cumulative variance plot in the 2D PCA plot.

The datapoints by PC1 and PC2 are represented by scatterplots in correlation. There are 12 unique clusters in the scatterplot. The places where the "Yes" and "No" values appear are shown by the blue and yellow circles.

Additionally presented are the findings regarding the correlations between the PCs and the other factors in the data set. The variables for age, marital status, and the dummy variable indicating that the patient is a child have the highest correlations in PC1. The factors relating to gender, smoking status, and heart disease have the largest relationships with PC2.

The locations of the points on the 2D PCA Plot are most strongly influenced by these variables.

The correlation between PCA and each variable reveals that the dummy variables for "male" and "female" entries have identical correlations with stroke but with opposing signs. In the end, this might invalidate the conclusions that can be drawn from the gender data. Future models should delve more deeply into how the variables that make up the PCs are composed.

Oversampling Data for Random Forest

The team must oversample the data in order to build reliable models for this set of information. Due to the significant class disparity between patients who had experienced a stroke and those who had not, this was necessary.

Oversampling's one drawback is the potential for overfitting the model. The output of the model is also slightly altered by oversampling the data. When using oversampled data, a model's variable importance will be different from when using the original data.

The group made the decision to develop a number of different models in order to examine the connections between the data set's variables and forecast who will have a stroke. We can better comprehend how these variables interact by examining the variable relationships. We can better understand how these variables increase or decrease a patient's risk of having a stroke by examining the relationships between the variables.

Random Forest Model - 97.4% Accuracy

The team ultimately decided to use the random forest model to forecast utilising the oversampled data. This model functions by assembling a number of decision trees into an ensemble. To create forecasts, our model consults 400 trees. This random forest model has strong specificity (96.3%) and sensitivity (100%) rates. The significance of the variables in terms of creating predictions was one of the insights that the team was able to gain from analysing the random forest model. We can shed some light on our first query, which was about what factors raise the risk of stroke, by determining which variables exhibit a higher degree of importance.

The next two most significant predictors of stroke are substantially more manageable because diet and exercise can control blood glucose levels and body mass index. Along with being aware of their importance, they can also be periodically examined to make sure that those who are battling with them are moving toward healthier levels. This makes them excellent preventative measures.

**STATISTICAL ANALYSIS**

The first glance at our data collection reveals 5,111 patient records. To describe the patient, each record has 11 numerical and category attributes. The following patient characteristics were gathered: the patient's gender, age, whether or not they had hypertension or heart disease, whether or not they had ever been married, what kind of work they do, what kind of neighbourhood they live in, their average glucose levels, their body mass index (BMI), whether they smoked, and whether or not they had experienced a stroke.

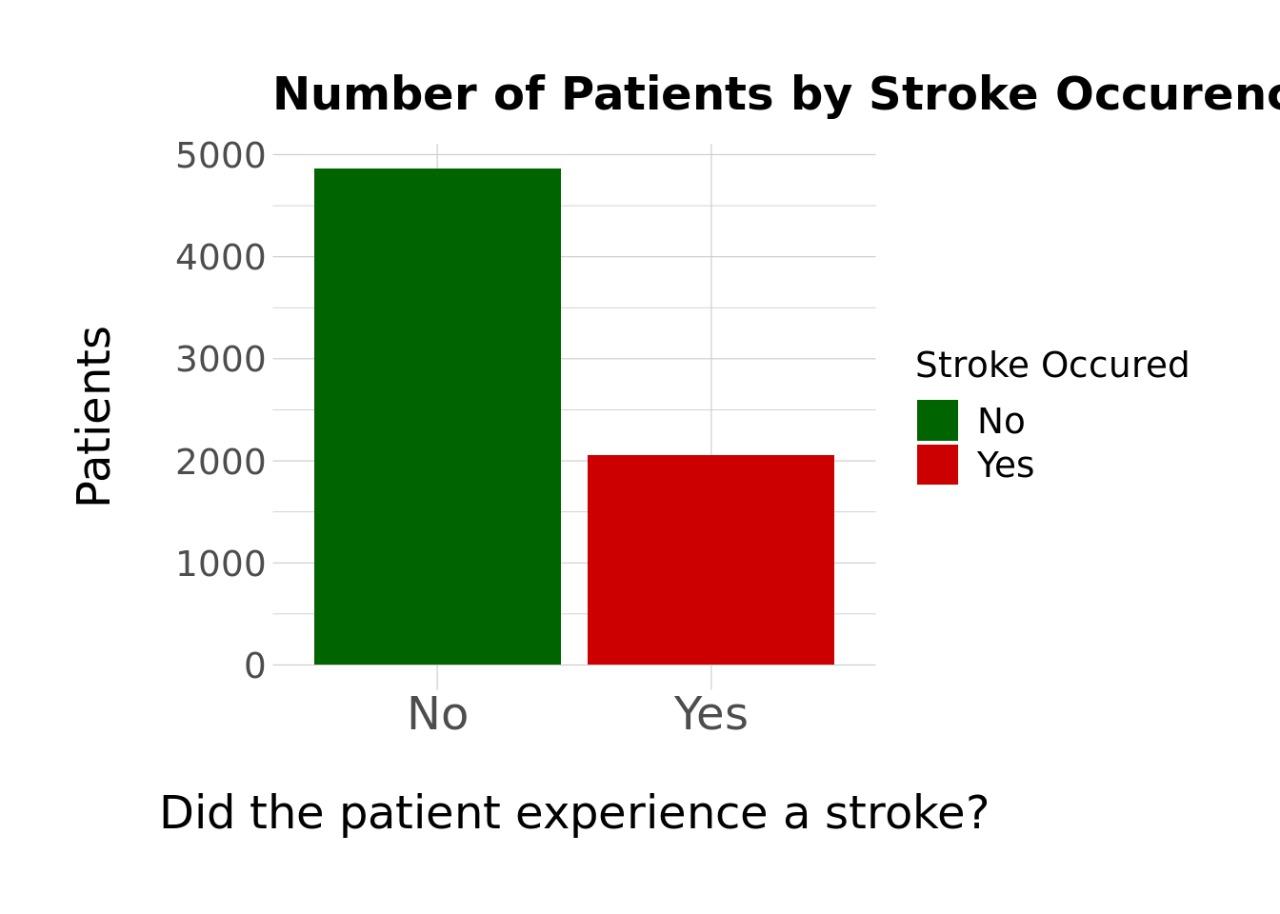
In the data set, there were 4,861 individuals without a stroke and 249 patients who had one out of all the fir entries. Patients' ages ranged from new-borns to 84. Of the patients, 2,994 were female, 2,115 were male, and 1 was classified as "other." The survey's average patient was 43.23, and it indicates that patient ages are somewhat regularly distributed.

With a mean age of 43.23 obtained from the summary() function, the study's patient population's ages exhibit a nearly normal distribution. The majority of patients are in their 40s, according to the data from the summary() function earlier and the chart above.

After adjustment, the summary() function's results for patient Body Mass Index is right-skewed, with a mean of 28.89. In Data Cleaning, all NAs were modified to mean.

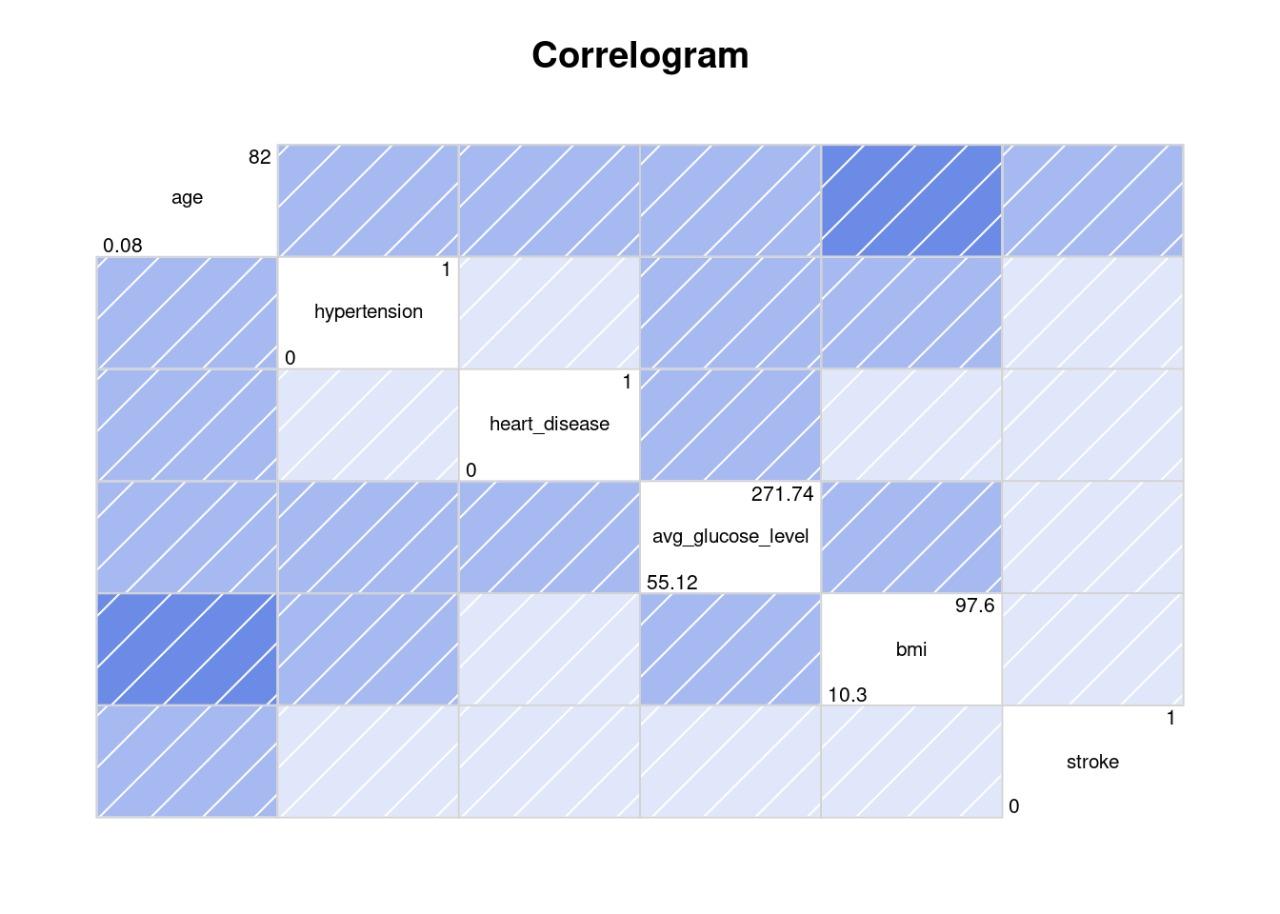
After adjustment, the summary() function's results for patient Body Mass Index is right-skewed, with a mean of 28.89. In Data Cleaning, all NAs were modified to mean.

Due to the fact that our data collection only consists of 5,110 patient records, the team chose to oversample the data rather than undersample it. Under sampling is frequently applied to data sets having 100,000 or more records. Oversampling does run the danger of overfitting the model, which is its one drawback. Additionally, slightly altering the data's oversampling can affect the model's output. A model that uses oversampled data will have a different variable importance than a model that uses the original data.



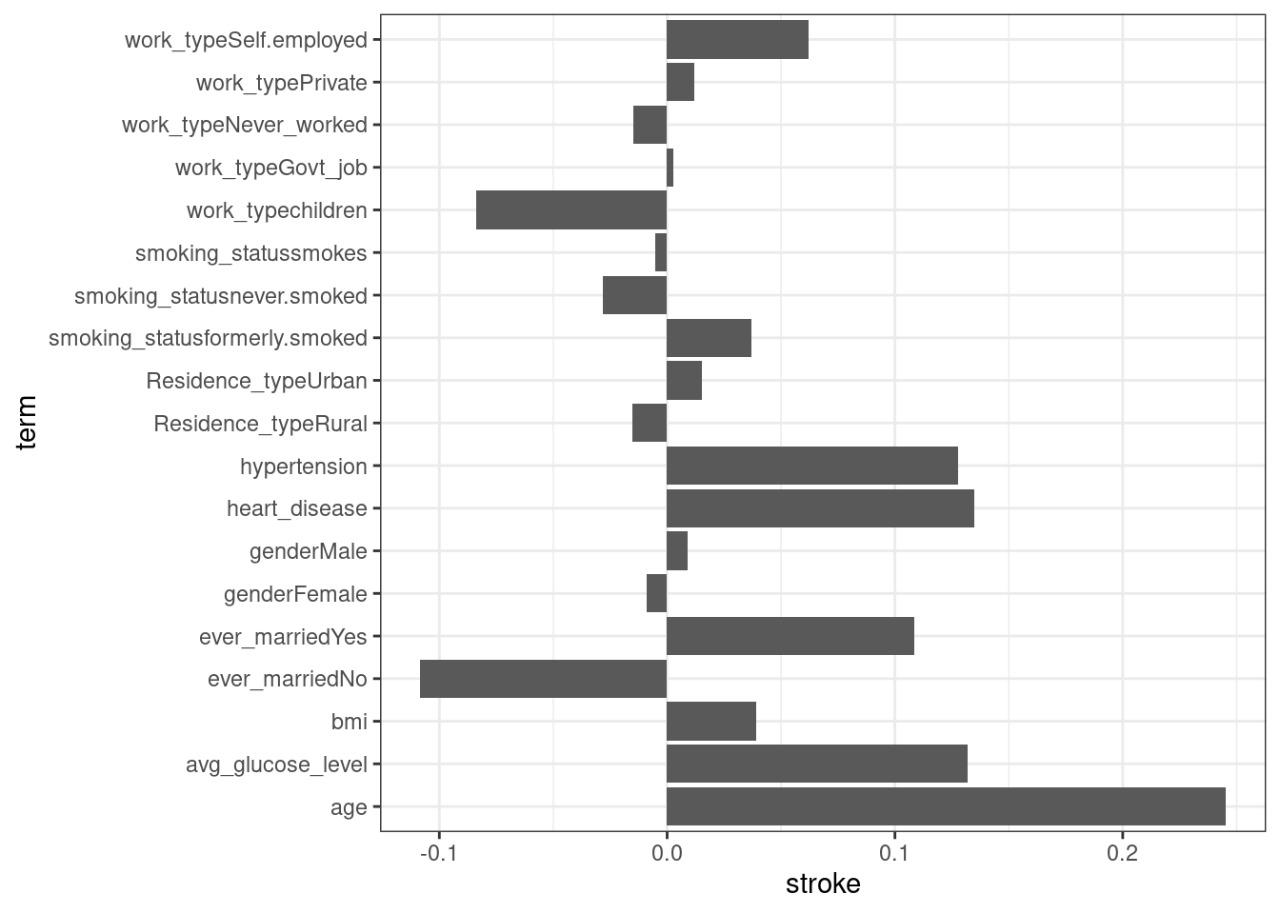
Correlogram

The correlation between each numerical variable in the cleaned data is displayed in the graph below. The minimum and maximum values are those found in the diagonal cells. As an illustration, the lowest bmi is 10.3 and the greatest bmi is 97.6. All numerical variables are favourably associated to the predictor variable [stroke], as shown by the correlogram and correlation table. The strongest association between age and stroke exists.

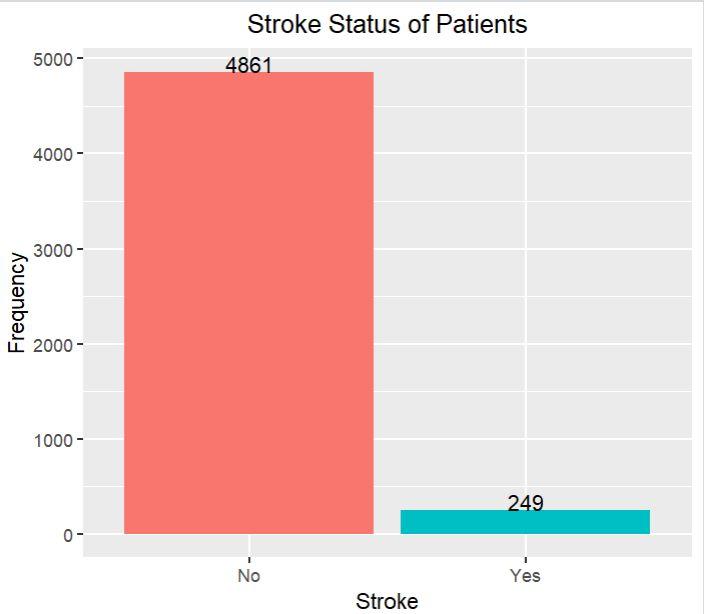


There are four main characteristics that influence whether a patient will receive a stroke diagnosis, according to the correlation matrix based on numerical variables. Age, hypertension, heart disease, and average glucose level are these factors.

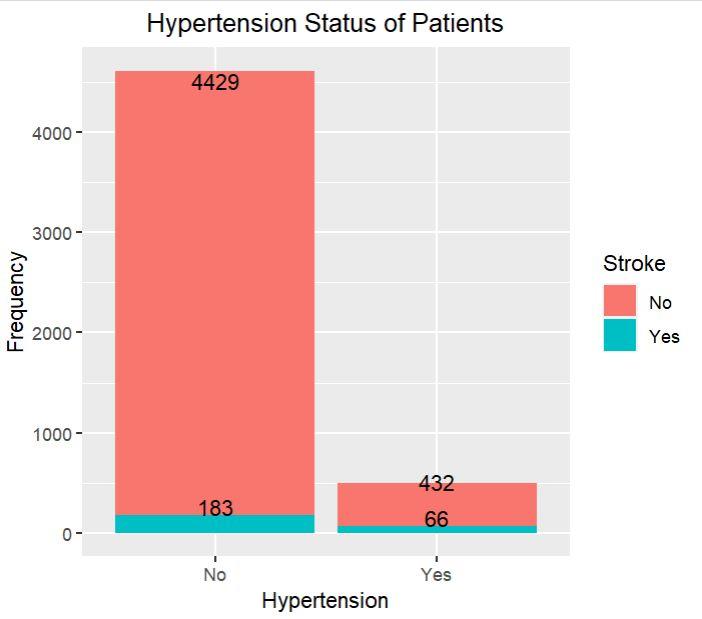
Numeral and categorical variables are both present in the plot below. The top 4 motivating factors are still the same as before. Of all the categorical variables, Ever married has the strongest correlation with stroke.



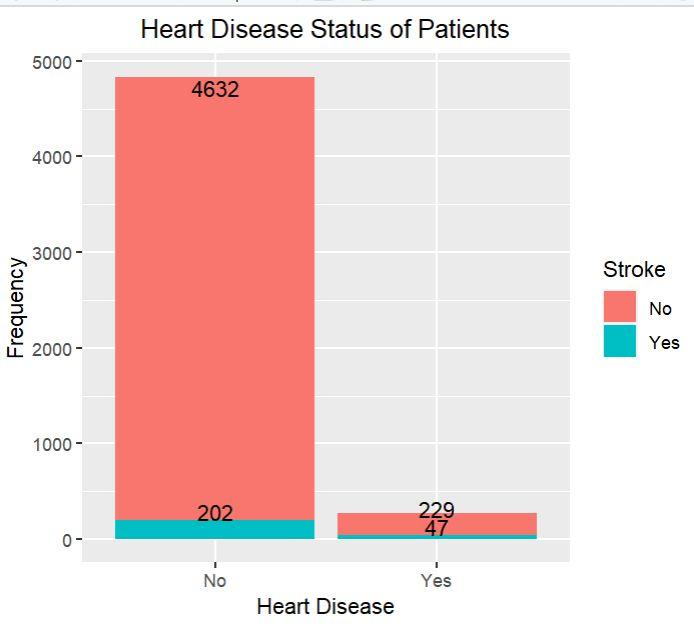
**VISUALIZATION ANALYSIS**



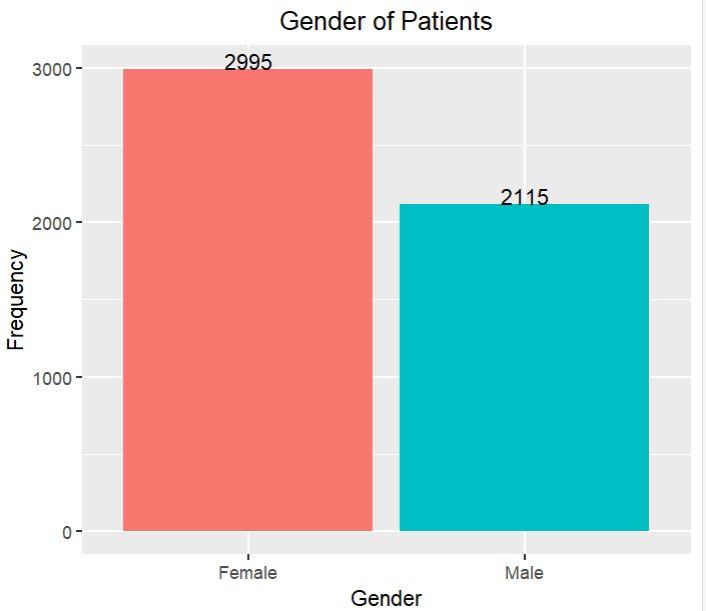
Patients who have not yet experienced a stroke outnumber those who have by a large margin.



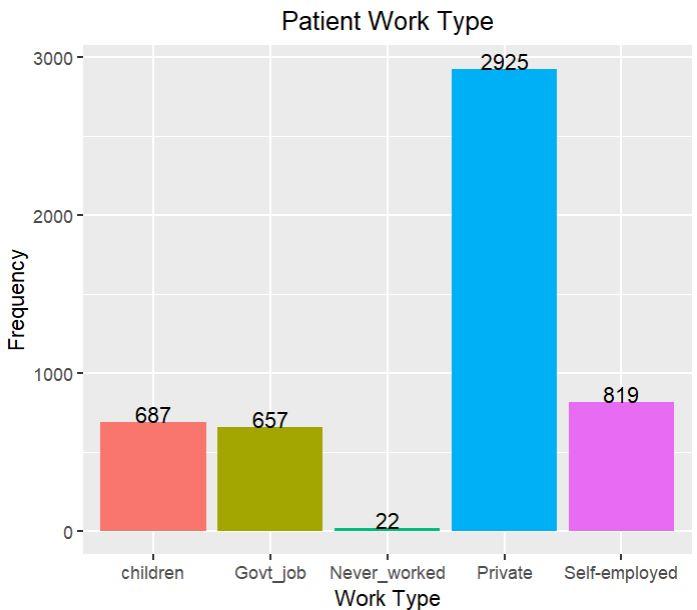
Although there are far more individuals without hypertension than those who do, the difference is slightly smaller than the difference between the two for stroke victims.



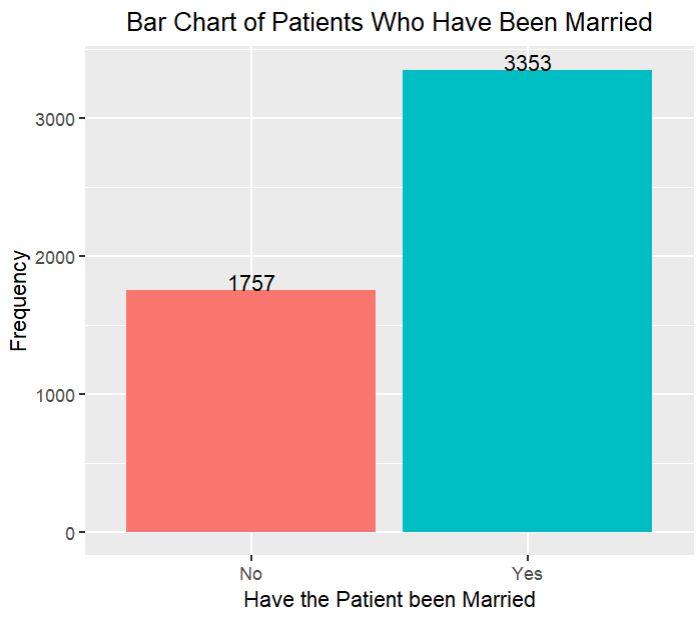
The difference between individuals with and without cardiac disease reflects the difference between people who have and have not had strokes more closely.



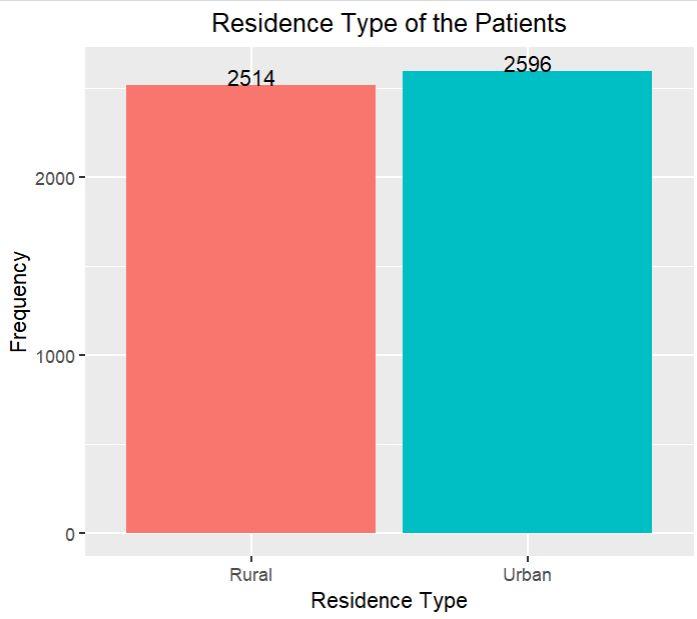
Patients are more likely to be women than men. Since the majority of the patients are female, the one entry that was marked as "Other" was added to the "Female" column.



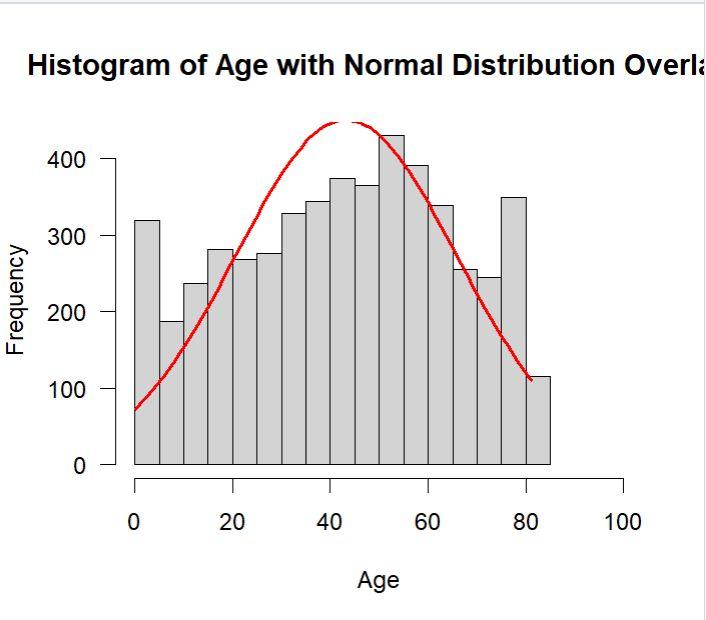
Patients who are self-employed, have children, or work for the government make up about an even number of patients. Most patients work for private businesses, albeit a tiny percentage never have.



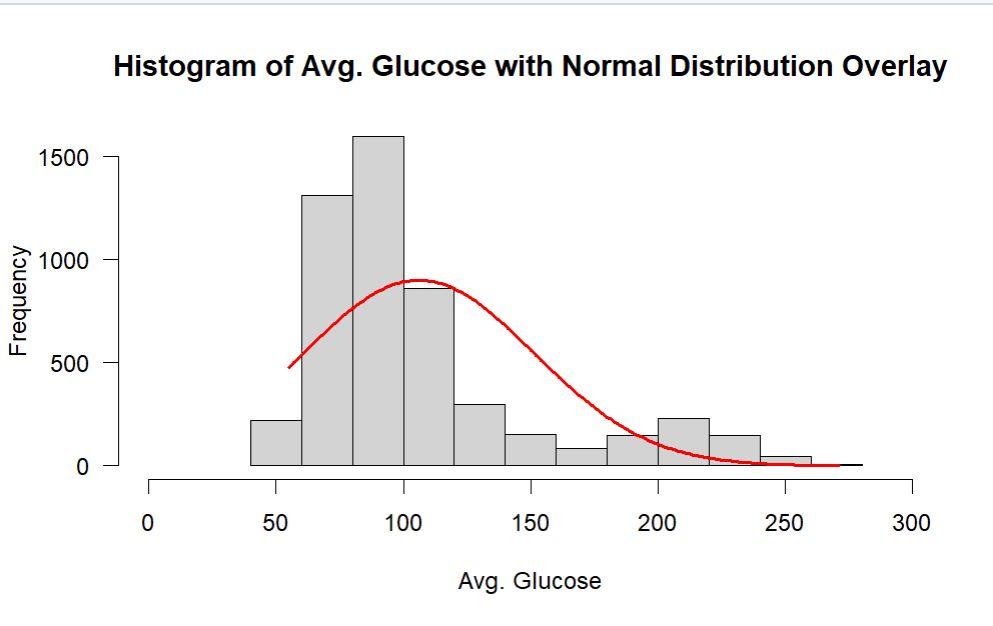
Patients who have previously been married are roughly twice as common as those who have not.



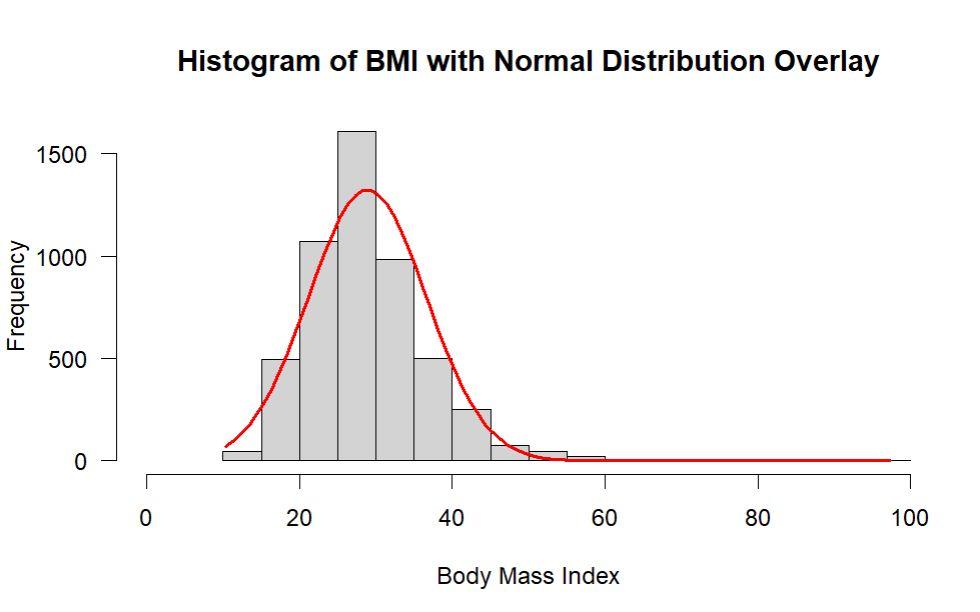
The patients are split almost equally between homes in rural and urban areas.



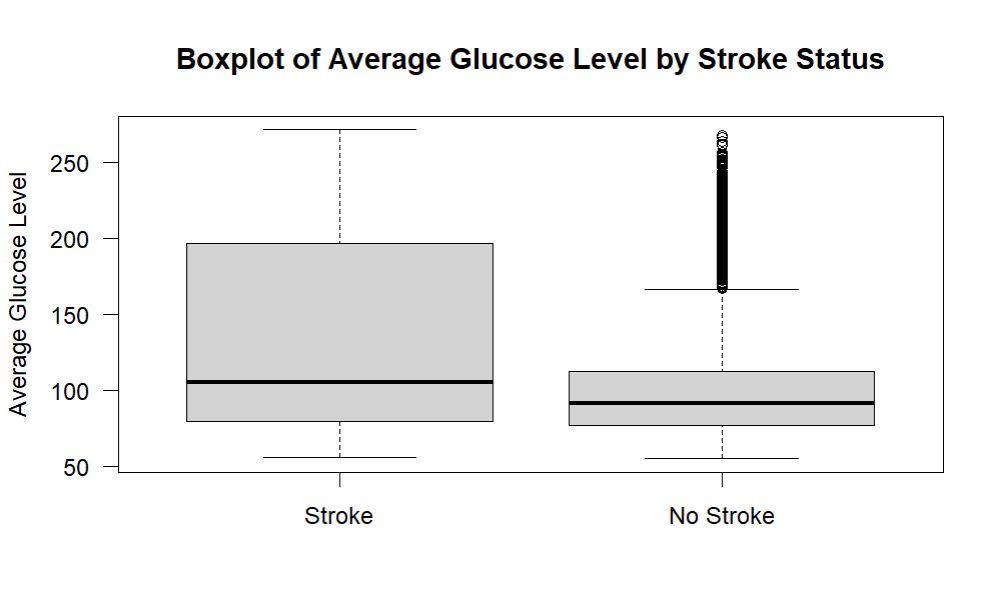
With a mean age of 43.23 obtained from the summary() function, the study's patient population's ages exhibit a nearly normal distribution. The majority of patients are in their 40s, according to the data from the summary() function earlier and the chart above.



With a mean age of 43.23 obtained from the summary() function, the study's patient population's ages exhibit a nearly normal distribution. The majority of patients are in their 40s, according to the data from the summary() function earlier and the chart above.



After adjustment, the summary() function's results for patient Body Mass Index is right-skewed, with a mean of 28.89. In the section on data cleaning's first part, all NAs were changed to mean.



The boxplot displays a mean Body Mass Index that is generally similar between patients who have had strokes and patients who have not, with a few high outliers in the former group and many high outliers in the latter.

**CONCLUSION**

In this project, we provided a targeted evaluation of patients’ attributes in digital fitness document for stroke prediction. We systematically analysed distinct functions. We achieved characteristic correlation evaluation and a step sensible evaluation for selecting an optimum set of functions. We discovered that the distinct functions aren't well-correlated. Additionally, we achieved important element evaluation. The evaluation confirmed that the majority important additives are had to give an explanation for a better variance. The variable loadings but confirmed that the primary important element which has the very best variance would possibly give an explanation for the underlying phenomenon of stroke prediction.

Finally, algorithms had been applied on a set of various functions and important additives configurations. The analysis and models were successful in finding the risk factors for strokes. The main takeaways of our analysis being that focusing on correct diet and exercise is big, especially as somebody gets older. Individuals who have a history of stroke in their family should be particularly careful to adhere and pay attention to guidelines around these risk factors. Anyone who has experience with any of these issues should speak with their primary care physician at next availability regarding stroke prevention. Additionally, there is some evidence to support that stress is related to strokes, and those in high stress environments should be aware of this risk and work to diminish their stress levels. Finally, our models indicated that individuals who suffer from Hypertension are additionally at risk and should be aware of this issue and speak with their physician about the risk from strokes. Hypertension was one of the variables that had high importance with several of the models that we ran, both with unsampled and original data. Those who are 50 years of age or older obviously cannot change their age, but they can focus on diminishing other risk factors such as hypertension, average glucose level, with methods such as healthy dieting, exercise, and including forms of stress reduction in their daily lives. The systematic evaluation of the unique capabilities withinside the electronic fitness data will help the clinicians in powerful archival of the data. Instead of recording and storing all the features, the data management done by us can archive only those features that are essential for stroke prediction.

Thus, in future, we plan to combine the digital information dataset with heritage expertise on one-of-a-kind sicknesses and pills the use of Semantic Web technology. Knowledge graph technology may be used with a view to put up the digital fitness information in an interoperable way to the studies community. The introduced heritage expertise from different datasets also can probably enhance the accuracy of stroke prediction fashions as well. We intend to acquire our institutional dataset for further benchmarking of those system studying strategies for stroke prediction. We also plan to carry out outside validation of our proposed method, as part of our upcoming deliberate work.