Prediction of Occurrence of Brain Stroke
ISM 6136- Data Mining Project

-Team Avengers
Sowmya Lakshmi Putta
Varshini Poreddy
Aiswarya Reddy Rangam
Sri Sai Teja Balusu
Ashwin Kumar Reddy Guduru

Health Care Analytics

Healthcare analytics is the practice of using data analysis to help healthcare companies improve outcomes, care quality, and operational effectiveness. It employs both historical and real-time data to generate actionable insights, enhance decision-making, and maximize results in the healthcare sector. Health outcomes and the patient experience can both be improved with the help of healthcare analytics, which has other advantages for healthcare companies as well. The healthcare sector is flooded with useful information in the form of detailed records. Healthcare providers are required by industry standards to keep many of these documents for a predetermined amount of time.

In order to anticipate patient demand and plan staffing, they use analytics techniques to forecast the availability of medical equipment (such as beds and ventilators). Analytics in healthcare might be useful for everyday patient care as well. Many businesses are utilizing linked technologies to give clinicians real-time data on a patient's critical health parameters as part of the expanding field of digital health in the healthcare industry. Chronic diseases like diabetes and high blood pressure can be better managed with the help of technology like remote patient monitoring and intelligent pill bottles and dispensers.

Background

In recent years strokes are one of the leading causes of death affecting the central nervous system. Stroke is rapidly developing clinical signs of focal (or global) disturbance of cerebral function, with symptoms lasting 24 hours or longer, or leading to death, with no apparent cause other than of vascular origin. Among different types of strokes, ischemic and hemorrhagic majorly damages the central nervous system. According to the World Health Organization (WHO), globally 3% of the population is affected by subarachnoid hemorrhage, 10% with intracerebral hemorrhage, and the majority of 87% with ischemic stroke.

Data from the recent INTERSTROKE study indicates that a list of ten stroke risk factors including hypertension, high cholesterol, current smoking, alcohol consumption, diabetes, stress, obesity, heart disease, lack of physical activity, and poor diet was responsible for 90% of all strokes.

1. **High Cholesterol:** Cholesterol or plaque build-up in the arteries can block normal blood flow to the brain and cause a stroke. High cholesterol may also increase your risk for stroke by raising your risk for brain stroke, a stroke risk factor.
2. **Alcohol consumption:** Drinking alcohol is related to the incidence of stroke. In general, the more alcohol that is consumed on the excessive side the greater the risk for the development of a stroke. This is true of both types of strokes (ischemic and hemorrhagic).
3. **Diabetes:** Diabetes increases the chance of having a stroke, which can damage brain tissue and cause disability or even death. Over time, excessive blood glucose can result in increased fatty deposits or clots in blood vessels. These clots can narrow or block blood vessels in the brain or neck, cutting off the blood supply, stopping oxygen from getting to the brain, and causing a stroke. To prevent stroke, people with diabetes should control blood glucose, blood pressure, cholesterol, and weight.
4. **Stress:** Stress can cause the heart to work harder, increase blood pressure, and increase sugar and fat levels in the blood. These things, in turn, can increase the risk of clots forming and traveling to the brain and heart, causing a stroke.
5. **Obesity:** Anyone who has put on a few too many pounds knows they can slow you down. Over time, if those pounds grow into obesity, they may do serious harm, putting you at risk for a wide range of illnesses. But too much weight on the body also can harm the brain.

6. **Lack of physical activity:** Being inactive can lead to fatty material building up in your arteries (the blood vessels that carry blood to your organs). If the arteries that carry blood to your heart get damaged and clogged, it can lead to a heart attack. If this happens in the arteries that carry blood to your brain it can lead to a stroke.
7. **Poor diet:** Red Meat. Eating steaks, sausages, and other red meats high in saturated fat could lead to a stroke. An elderly person who eats red meat is more likely to have an ischemic stroke because saturated fat causes blockages in blood vessels that supply blood to the brain. Up to 80 percent could be avoided stroke with healthy lifestyle changes, including eating better, exercising regularly, quitting smoking, and losing weight if you're overweight or obese.

These risk factors, which are predominantly traditional risk factors of stroke, are modifiable, making stroke highly preventable. Knowing your stroke risk factors, following your health care provider's recommendations and adopting a healthy lifestyle are the best steps you can take to prevent a stroke. If you've had a stroke or a transient ischemic attack (TIA), these measures might help prevent another stroke. The follow-up care you receive in the hospital and afterward also may play a role.

Problem Statement

The overall objective will be to predict the occurrence of brain stroke accurately with few tests and attributes. Attributes considered form the primary basis for tests and give accurate results. Many more attributes can be taken but our goal is to predict with few attributes and faster efficiency, the risk of getting a brain stroke. This practice leads to unwanted biases, errors, and excessive medical costs which effect the quality of service provided to patients. So our main two problem statements will be

1. Do we need to do this so early in life and worry about it all long?
2. Will we be able to save the treatment costs if knowing the occurrence of brain stroke in the early stages?

Solution Methodology

We used the Kaggle repository dataset for this investigation. The dataset has 11 properties, 8 of which have been filtered and are frequently utilized for study. We would need to perform a ROC (Receiver Operating Characteristic) analysis on the distribution of true positive rates and false positive rates as a starting point for developing evaluation criteria. Additionally, we would be comparing the effectiveness of different categorization algorithms using criteria like Precision, Accuracy, Recall, and F1 Score. Two class Neural Networks, Two class Logistic Regression, Two class Boosted Decision Tree, and two class Decision forest were the four algorithms used in the implementation of this classification solution. The dataset was completely divided into training and testing datasets in a ratio of 70% to 30%. Then we looked at the AUC-ROC curve, which shows TPR (Sensitivity) vs. FPR (1-Specificity) at various classification thresholds, as one of our evaluation measures. More items will be categorized as positive as the classification threshold is dropped, leading to an increase in both False Positives and True Positives.

Greater coverage of the curve's underside by a more accurate model results in a higher ratio of true positives to false positives overall. Out of the 4 machine learning methods, the findings showed that two class decision forest covered a larger area than neural network, decision tree and logistic regression models, demonstrating the model's effectiveness. The Confusion Matrix is another often employed evaluation metric for categorization issues. High True Positives and True Negatives but

low False Positives and False Negatives are characteristics of the ideal model. This metric would also enable us to determine which algorithm performs better than the other.

Dataset Description: The dataset consists of 4981 instances and 11 attributes. Out of which 10 are independent variables and one is the dependent variable.

Independent Variables
Hypertension: Whether the patient has high blood pressure or not. (1 if the patient has hypertension and 0 if does not have hypertension)
Heart Disease: Whether the patient has heart disease or not. (1 if the patient has heart disease and 0 if he/she does not have heart disease)
Ever Married: If the patient is married or not.
Residence Type: If the patient lives in a rural or urban area
Work Type: It describes the type of work whether the patient does a private job, govt. job, or if the patient is self-employed and if the patient is a child.
Average Glucose Level: Average Glucose level(mg/dl)
Smoking: Whether the patient has a habit of smoking previously, currently, unknown, or never smoked.
BMI: Body Mass Index. Weight in kilograms/height meters squared.
Age: Patient age in years. Age is a risk factor, in other words, the higher the age, the more likely that the patient is at risk of having a brain stroke
Gender: Whether the patient is a male or female.

Dependent Variable
Stroke: It's the output which tells the chance of having a brain stroke (Value=0 less than 50% chance of getting a brain stroke, Value=1 more than 50% chance of getting a brain stroke).

Algorithm Comparison:
The target variable, which after dataset analysis indicates is of discrete type and predicts the possibility of occurrence of brain stroke in patients, suggests that the type of data mining problem we are dealing with is one of binary classification. We will thus take into consideration the most popular and efficient Binary classification methods, which are listed below, in order to develop a data mining solution to this problem and to estimate this likelihood.

Two-Class Decision Forest
This decision forest algorithm is an ensemble learning method intended for classification tasks. Ensemble methods are based on the general principle that rather than relying on a single model, you can get better results and a more generalized model by creating multiple related models and combining them in some way.

Two-Class Logistic Regression
Logistic regression is designed for two-class problems, modeling the target using a binomial probability distribution function. Two-Class Logistic Regression module to create a logistic regression model that can be used to predict two (and only two) outcomes. Logistic regression is a well-known statistical technique that is used for modeling any kind of problem.

Two-Class Neural Network

A neural network is a set of interconnected layers. The inputs are the first layer and are connected to an output layer by an acyclic graph comprised of weighted edges and nodes. Classification using neural networks is a supervised learning method, and therefore requires a tagged dataset, which includes a label column.

Two-Class Boosted Decision Tree

A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction.

Preprocessing and Data Cleaning

The dataset that we took is first imported and then we used the summarize data feature to check is there are any missing values and check the data type of each attribute. As the gender and residence type attributes do not contribute much to the dependent variable, we remove these attributes by using select columns in the dataset feature. We then used the edit metadata feature to convert the string type attributes to categorical datatypes. To bring all the numerical data to a common scale of -1 to 1 we use the normalized data feature.

Then we encode the data using the group categorical feature where we label encoded the ever-married attribute which has two unique values in numeric form/machine-readable form (0 and 1). In a similar way we label encoded work type and smoking status which has four unique values each converted into machine readable form (i.e., 0,1,2,3). This preprocessed data is then given as input to the split data feature.

Final Experiment & Models:

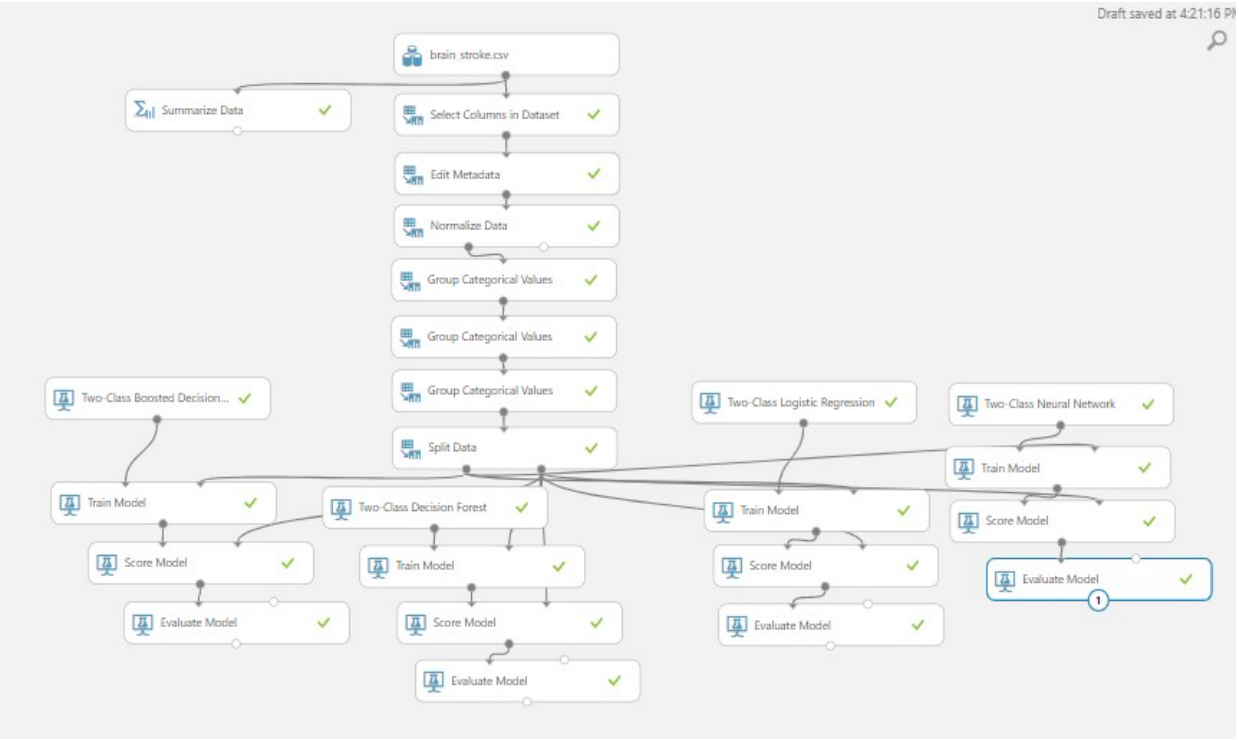

Figure 1: Azure ML studio Brain stroke data model

Split data: This will divide the dataset into train dataset and test dataset and for this experiment we
took 70 percent as train data and 30 percent as test data.

Train Model: The training set from the split data is given as input to the train model with the corresponding algorithm. We have trained the data with four different algorithms which are a two-class boosted decision tree, two-class decision forest, two-class logistic regression and two-class neural network. The output of the train model is given as input to the score model

Score Model: Score Model outputs a predicted value for the class, as well as the probability of the predicted value. Here we give the output of the train data and test data as the input.

Evaluation Model: We use it to evaluate our scored classification by using standard
metrics and we can visualize all parameters like ROC, AUC, Recall, precision, and F1 score
from the evaluation model.
Now that we have identified the dataset and a utilized couple of classification models over the data, let us look at the results obtained.

Results
Think about the case where a patient is experiencing a brain stroke, but the model was unable to foresee it. Such scenarios are undesirable, and we don't want our model to anticipate patient's brain strokes under such circumstances. We want to have as few False Negative numbers as possible based on the Confusion matrix. Recall value is inversely proportional to False Negatives, which means that the higher the recall value, the more accurate the model is in this situation. Additionally, we must consider the possibility that the projected person would get a brain stroke, which will reveal information about precision.
In our case, false negative i.e., the person had a stroke, but the model predicted that the person didn't have a stroke has a lot of impacts. In such cases recall is considered the best evaluation metric as we need to have fewer false negatives for our data. Below are the evaluation metrics that are to be considered for estimating the best algorithm for our dataset.

Figure 2. Two Class Boosted Decision Tree

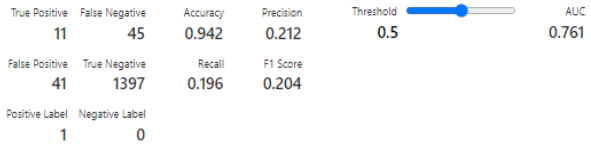

Figure 3. Two Class Logistic Regression Model

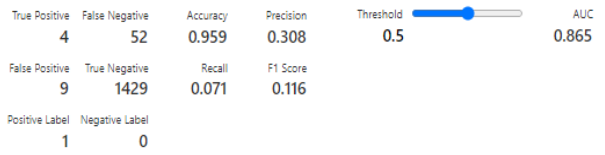

Figure 4.Two class Neural Network model

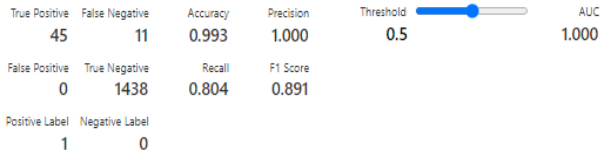

Figure 5. Two class Decision Forest model

ROC CURVE:

Looking at the ROC Curve from Figures 2 and 3. To determine whether it is good or bad can be identified using Area Under the Curve (AUC) metrics. More the Area Under the Curve, the more the model can explain the variation in the data. Hence, greater the AUC value the best fit the model is for the available data.

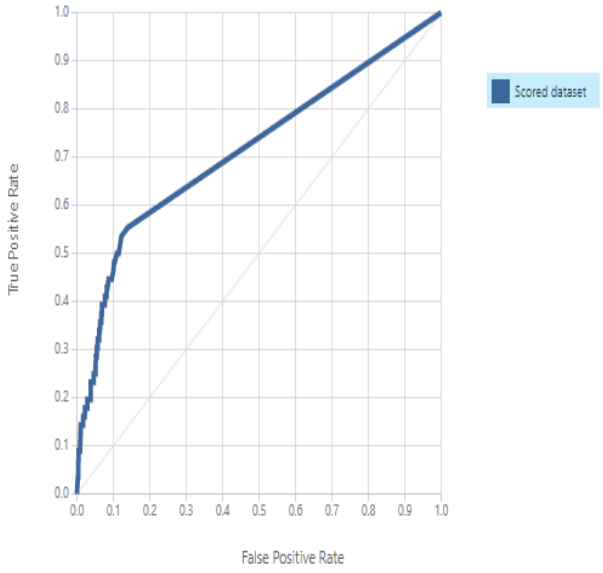

Figure 6. Two Class Boosted Decision Tree

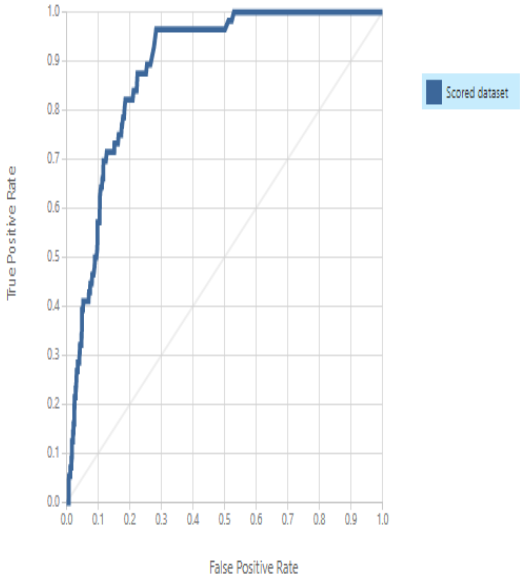

Figure 7. Two Class Logistic Regression Model

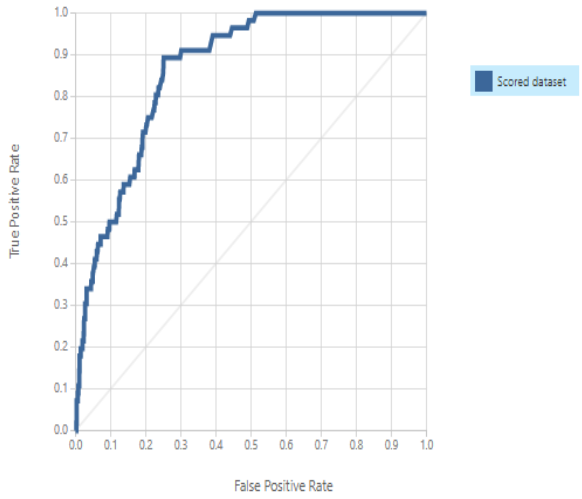

Figure 8.Two class Neural Network model

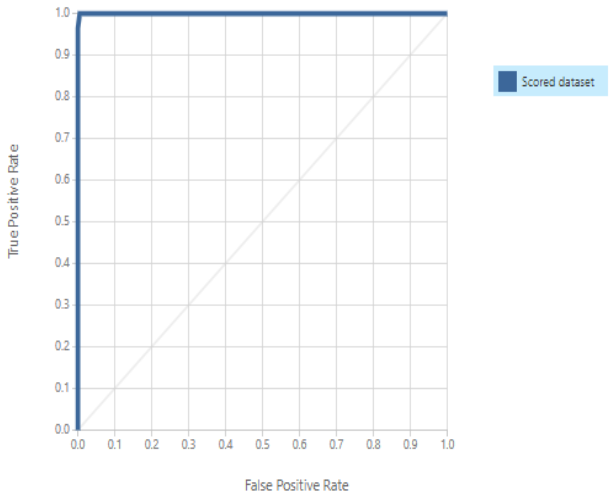

Figure 9. Two class Decision Forest model

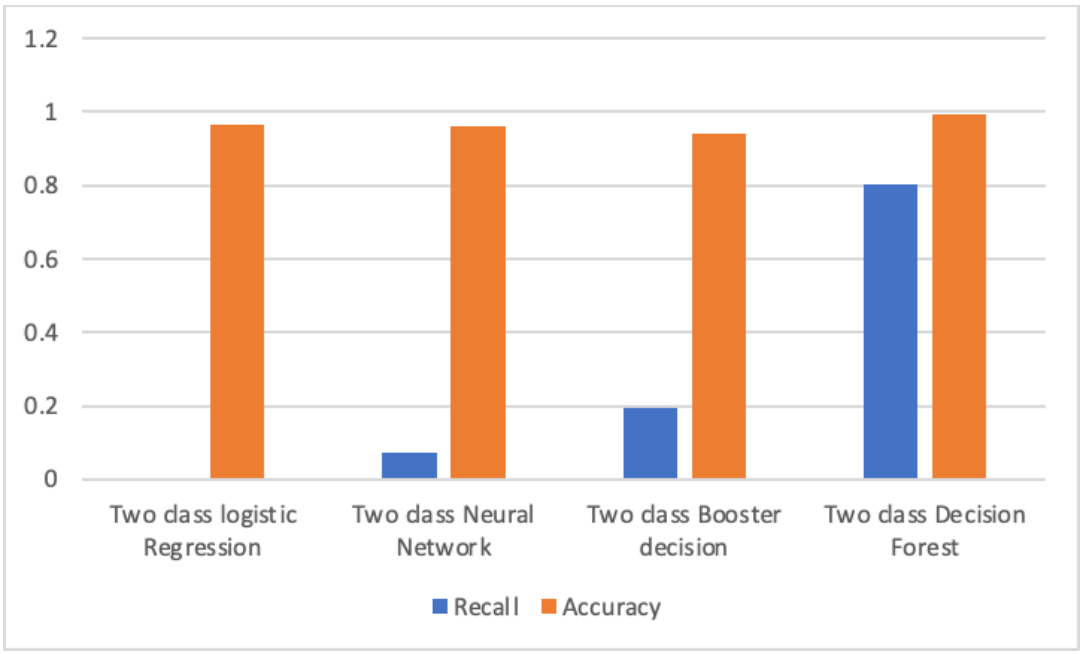

Figure 10. Evaluation for different models

From figure 10 by observation all the algorithms have good accuracy but as we consider recall value for our dataset we can see that the recall value is highest i.e., 80% for two class decision forest. By this we can conclude that the two class decision forest is the best of the above models.

Conclusion:

One can wonder if everyone should learn about their health problem so early in life and worry about it constantly until they are cured. The answer is that it would be preferable to be aware of their health status and take appropriate action as today's youth spend a lot of time, effort, and money today in order to have better results tomorrow, whether it be in education, cryptocurrencies,

stock markets, real estate, etc. Nowadays, ministrokes also signifies that the person may get stroke sooner or later. Early detection of signals can certainly assist the physicians to start quick treatments for the prevention of strokes. The damage of stroke is irreversible, and prevention is the only cure for strokes. Due to covid-19 people have started investing in healthcare and have become very conscious about how they monitor their health. Wearables are also coming up with new sensors and advanced features which can monitor their health conditions and in recent times new phone applications have been developed which analyze the data from the wearables. Data prediction techniques in stroke imaging could markedly change the milieu of stroke diagnosis and management in the near future. Prediction ability to provide clinically relevant output information solely depends on the correctness of the input data and the machine learning method used to train the AI model. Yet, the timely diagnosis of stroke is vital for functional recovery and to minimize mortality. In addition, foreseeing the degree of post-stroke recovery in advance and informing patients/family members of the prognosis may enhance the treatment rapport and rehabilitation processes. New advancements in imaging technology for stroke diagnosis have led to the availability of a large volume of scattered neuroimaging information. Here, artificial intelligence and machine learning have been employed in several ways to extract the most coherent information, which can be used as an identifier or marker for stroke diagnosis and for analyzing its severity. this valuable data should not fall into the hands of Money Mongers such as Big Private Hospitals and Health Insurance Companies. We all know what they can do with such kind of valuable data. So, to solve this situation the data should be centralized and be kept with the Government, so that they can identify the individuals who are prone to have brain stroke and act accordingly for better lives of their citizens.