HeathCare: Stroke Prediction Problem

Abhishek Kumar\*, Dr. Abdul Serwadda  
Department of Computer Science, Texas Tech University, Lubbock, Texas 79409, USA

**Abstract**— A stroke occurs when parts of the brain suffer from insufficient oxygen supply. Stroke is the fifth leading cause of death in the United States. The death toll in the United States is 140,000 each year. 1 of every 20 deaths is the result of stroke. The healthcare dataset stroke data provided by the official kaggle website is choosen for our research purposes. We will analyze various attributes from our dataset and brainstorm to find out which ones are the most appropriate for our purpose. Initially, we perform basic data cleaning and data preprocessing to handle the missing datas. Further, we explore our dataset by performing outlier detection and plotting distributions to study about our data. Moreover, feature engineering is performed to find the interactions between the various features of our data. We try various methods of dimensionality reduction to further analyze our features. A list of standard machine learning algorithms such as decision tree, random forest, logistic regression and XGBoost Classifier were used for our model building. Our stroke dataset is a scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes. This comes under the imbalance class distribution. To resolve this, we balanced our dataset and calculated results with and without data preprocessing. The best accuracy for our dataset was when we used a balanced dataset with feature selection.

**Index Terms**—stroke prediction, machine learning, deep learning, imbalance dataset, pattern recognition, data preprocessing, XGBoost classifier

**GitHub Link** - <https://github.com/abhishek-kumar-code/HealthCare-StokePrediction>

Printed in IEEE Computer Society Standards

—————————— ◆ ——————————

# 1 Introduction

O

VER 15 million people suffer from stroke worldwide every year. Of these, 5 million die and 5 million are permanently disabled. A stroke occurs when parts of the brain suffer from insufficient oxygen supply due to either a blood clot (ischaemic stroke, IS), or a ruptured vessel causing an intracranial haemorrhage (ICH). It is estimated that in year 2030, stroke will be one of the four leading causes of death. However, accurate prediction of stroke is highly valuable for early intervention and treatment. Also, if the treatment can be given quickly after stroke, the chances of permanent disability reduces to a great extent.

Every 4 minutes someone dies from stroke. Up to 80 percent of strokes can be prevented. How a person is affected by their stroke depends on where the stroke occurs in the brain and how much the brain is damaged. For example, someone who had a small stroke may only have minor problems such as temporary weakness of an arm or leg. People who have larger strokes may be permanently paralyzed on one side of their body or lose their ability to speak. Some people recover completely from strokes, but more than 2/3 of survivors will have some type of disability.

————————————————

1. Abhishek Kumar is with the Department of Computer Science, Texas Tech University, Lubbock, Texas 79409, USA. E-mail: abhishek.kumar@ttu.edu.
2. Dr. Abdul Serwadda is with the the Department of Computer Science, Texas Tech University, Lubbock, Texas 79409, USA. E-mail: abdul.serwedda@ttu.edu.

**\*\*\*Please provide a complete mailing address for each author, as this is the address the 10 complimentary reprints of your paper will be sent**

*Please note that this work is part of the CS 5341 Pattern Recognition course conducted in Fall 2018 by the Depatment of Comuter Science, Texas Tech University.*

Fig. 1. Top 10 global causes of death in 2016

In this project we are using machine learning techniques to

predict stroke based on various health aspects of the patient.

# 2 Healthcare Dataset Stroke Data

## 2.1 Introduction to the dataset

The healthcare dataset that we have used in our research can be found on the official website of Kaggle, URL: <https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data>. The data source is the file named train\_2v.csv that contains 43,400 rows and 11 columns. The data type of this column varies from string value to integer and decimal. Figure 2 shows all the features of our dataset. Each patient is identified uniquely by a 5 digit ‘id’. The ‘gender’ of the patient is a string of value male or female. Other string value features include ‘ever\_married’, ‘work\_type’, ‘residence\_type’ and ‘smoking\_status’. Features like ‘hypertension’, ‘heart\_disease’ and ‘stroke’ are repreneted by binary values. ‘Age’, ‘avg-glucose\_level’ and ‘bmi’ are registered in decimal values.

Fig. 2. Definitions of the columns of the dataset

## 2.2 Dataset Analysis

From the last column of our dataset we can extract the stroke dataframe which suggests that we have 783 stroke pateints amongst 43,400 values i.e., approximately 1.804% of our patients are suffering from stroke. Looking at our dataset from a different frame of reference tells us that the work type category of *Private jobs* have the highest number of counts of stroke patients. This is around 1.75 times more than *self-emplyed people* and around 5 times more than employees working for the *government*.

Our data set can be broadly classified under two categories of male and female gender classes. We have 59.13% of females in our dataset and 1.67% of the total female patients have suffered from stroke. On the other hand, males have the tendency of suffering from stroke more than the females do. Our dataset comprises of 40.83% of males and almost 2% of these have a chances of getting stroke. These data anlysis have been shown in Fig 3.

|  |  |
| --- | --- |
| Stroke | Count |
| 1 | 783 |
| 0 | 42,617 |

|  |  |
| --- | --- |
| **Work Type** | **Work Type Count** |
| Private | 441 |
| Self-Employed | 251 |
| Govt-Job | 89 |
| Children | 2 |

|  |  |  |
| --- | --- | --- |
| **Gender** | **Gender Count** | **Percent** |
| Female | 25,665 | 59.1359 |
| Male | 17,724 | 40.8387 |
| Other | 11 | 0.0253 |

|  |  |  |
| --- | --- | --- |
| **Gender** | **Gender Count** | **Percent** |
| Male | 352 | 1.9860 |
| Female | 431 | 1.6793 |

Fig. 3. Tablewise data analysis of people having stroke

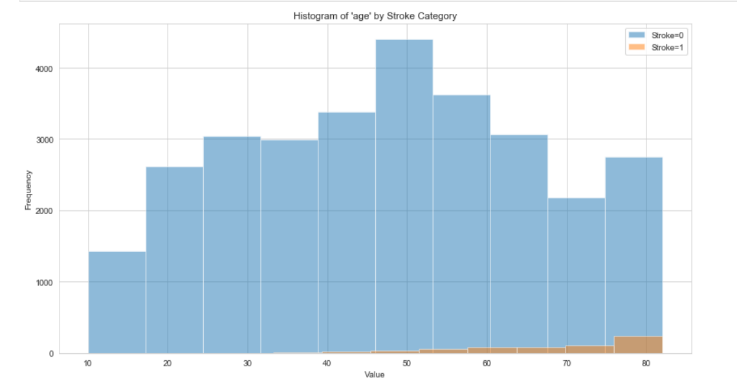
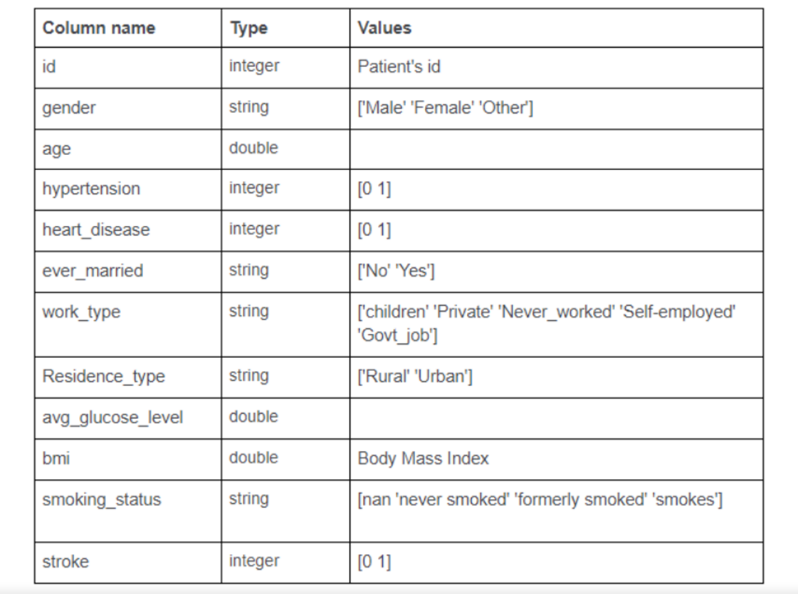
Analysis of our data led to the fact that age has an influence and is an important risk factor for stroke. Fig 4. is a visual representation of how age affects one’s chances of having stroke.

Fig 4. Histogram of age by stroke category.

Other features like ‘hypertension’, ‘ever\_married’ and ‘heart\_disease’ were not the best to make predictions as they were binary value representation. To realize the importance of other features and how they behaved we performed feature selection using SelectKBest method in Python that is further discussed in the Feature Selection section.

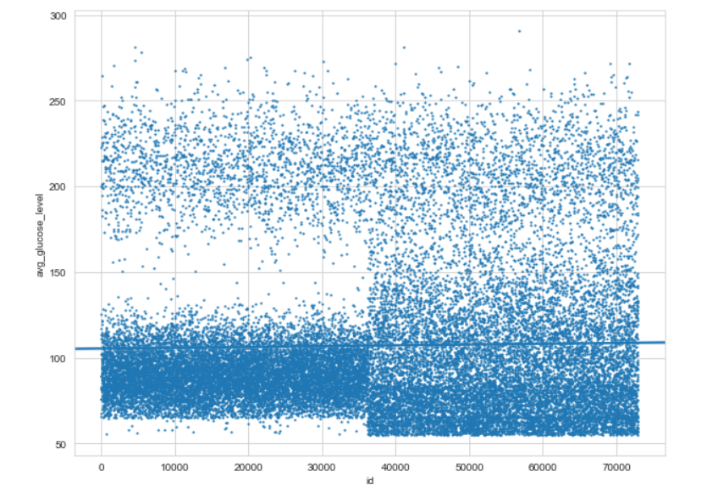
## 2.3 Data Cleaning

From a total record of 43,400 patients, we got a total of 13,292 missing values in the category of smoking status. The ‘bmi’ category had 1462 missing values. Since the frequency of missing values for bmi were low, we used it’s mean value (~28.6) to replace them. A huge percentage (~30%) of smoking imformation was not available so we treated them as 2 separate data tables – one with removing the rows for missing values and another with encoding. The patient id column was removed to prevent overfitting. The reason behind this is that a powerful classifier would use that column to fit perfectly on the training set, ignoring all the other columns, resulting in a model that is useless.

From the 11 columns, we had categorical type in 5 of them. One hot encoder was used to perform *binarization* of the category and include it as a feature to train the model. None of the categories from any column were removed before encoding as all of them had significant number of data. As a result, we ended up with 21 columns of data.

## 2.4 Data Exploration – Outlier Detection

We tried to study our outliers using the histogram and the scatter plots to see a general view of how the values vary from the normal. The two features that we were most interested in were ‘bmi’ and ‘avg\_glucose\_level’. Fig. 3 and Fig. 4 show the variation of data for there respective units. A horizontal line passing through the data points shows where the majority of the data lies.

Fig 5. Scatter Plot to uncover ‘bmi’ anomalies

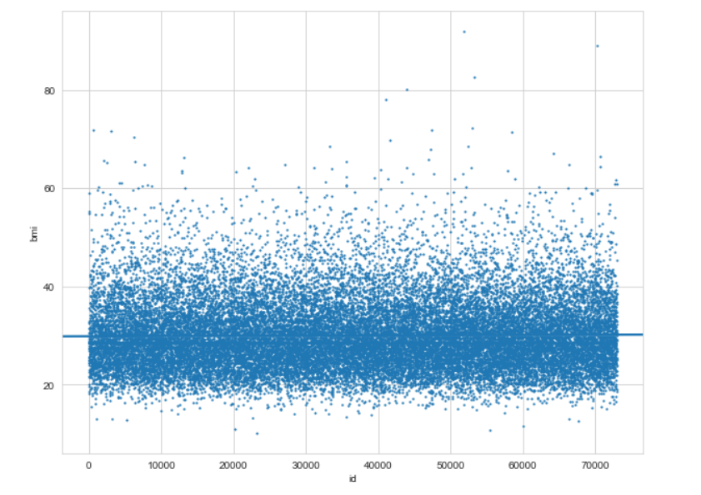


Fig 6. Scatter Plot to uncover ‘avg\_glucose\_level’ anomalies

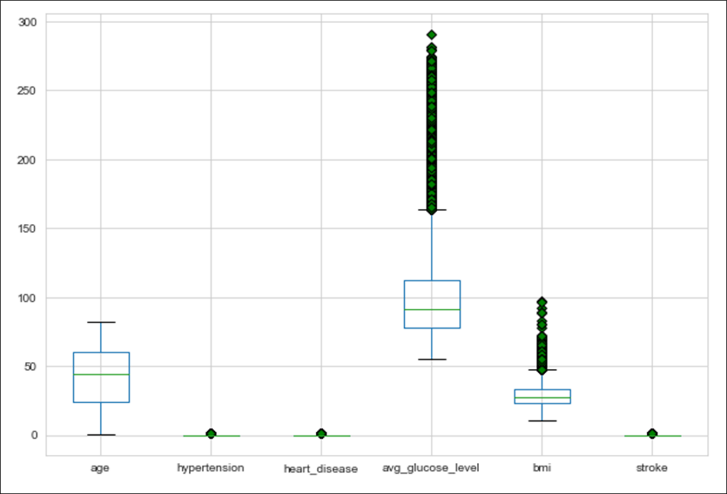
There are no special techniques within the scatter plot that tell us that the points are outliers. One has to use his/her intuition to make any analysis.

BMI is used to broadly define different weight groups in adults with the same groups applying to both men and women. The following categories makes it easy to undesratand: (i) Underweight: BMI is less than 18.5, (ii) Normal weight: BMI is 18.5 to 24.9, (iii) Overweight: BMI is 25 to 29.9, (iv) Obese: BMI is 30 or more.

A normal fasting (no food for eight hours) blood sugar level is between 70 and 99 mg/dL. A normal blood sugar level two hours after eating is less than 140 mg/dL

The better way is to analyze it using advanced statistics. One of the approaches choosen for our research purposes include Tukey Outlier Detection Method.

When there are no outliers in a sample, the mean and standard deviation are used to summarize a typical value and the variability in the sample, respectively. When there are outliers in a sample, the median and interquartile range are used to summarize a typical value and the variability in the sample, respectively. Tukey Method is based on the following: “Outliers are values below Q1-1.5(Q3-Q1) or above Q3+1.5(Q3-Q1) or equivalently, values below Q1-1.5 IQR or above Q3+1.5 IQR.”

Fig 7. Outlier detection of our features using Tukey Box Plots

## Imbalace Dataset

If you have spent some time in machine learning and data science, you would have definitely come across imbalanced class distribution. This is a scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes.

This problem is predominant in scenarios where anomaly detection is crucial like electricity pilferage, fraudulent transactions in banks, identification of rare diseases, etc. In this situation, the predictive model developed using conventional machine learning algorithms could be biased and inaccurate. This happens because Machine Learning Algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not take into account the class distribution or balance of classes.

# 3 Feature Engineering

In the set of surveys answered by each patient, there are several features of contiguous questions designed to evaluate a similar quality. For instance, bmi, avg\_glucose\_level, age, and hypertension are somewhere correlated. Therefore we looked through the descriptions for each feature and decided to aggregate a group of features.

A simple two-way interaction is represented by: X3 = X1 \* X2, where X3 is the interaction between X1 and X2. We will calculate two-way interactions among all features. After feature combination among all features, we ended up with 207 columns. As this might be computationally expensive, we will do feature selection after this. We performed feature selection using SelectKBest with k as 20.

# 4 Modelling And Performance

## 4.1 Algorithms

We used the train test split method to split training and testing data into 80% and 20%. The total data after preprocessing was: 30,108 from 43,400

The selected algorithms for classification include Decision Tree, Random Forest, Logistic Regression, XGBoost classifier, Application after balancing dataset.

## 6.2 Evaluation and Training

The first experiment we performed was without feature selection using the dataframe that contains all removed rows for missing values of smoking status. The models used for that were Logistic Regression, Decision Tree and Random Forest classifier. We calculated the number of samples for 1%, 10%, and 100% of the training data. The second experiment was done with and without feature selection and without deleting the rows for smoking status missing values.

# 7 Results

## 7.1 For Imbalanced dataset

We used several algorithms as mentioned above and got the results as shown in Fig 4. The training time was lowest for Decision Tree and Highest for Random Forest Classifier. The training accuracy was found to be same for Decision Tree and Random Forest which was 99%. The testing accuracy was higher in Logistic Regression and Random Forest which was 97%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic Regression** | **Decision Tree** | **Random Forest** |
| Train Time | 16 sec | 14 sec | 24 sec |
| Training accuracy | 98% | 99% | 99% |
| Testing accuracy | 97% | 94% | 97% |
| F-train | 98% | 99% | 99% |
| F-test | 97% | 94% | 97% |

Fig. 8 Experimental analysis result

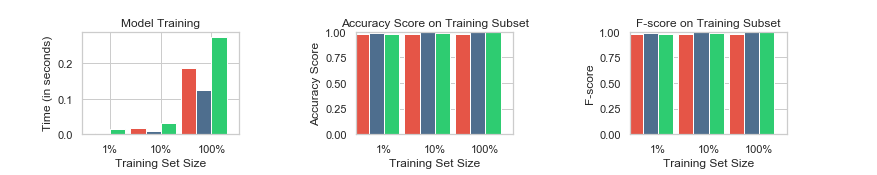
Figure 9 shows the performance metrics for 3 supervised learning models. Random forest classifier showed to have very high accuracy for both training and testing of 99%. Although the accuracy seems to be very high, it is clearly misleading because of the imbalanced dataset. Since we have maximum of patients without stroke our model seems to be working well.

Fig. 9 Performance metrics for supervised learning models

|  |  |  |  |
| --- | --- | --- | --- |
| **Model used** | **AUC without preprocessing** | **AUC without preprocessing** | **Model improvement with preprocessing** |
| Logistic Regression | 0.781863 | 0.695055 | 12.489% |

Fig. 10 Results for modeling with feature engineering

## 7.2 For balanced dataset

Finally, we tried making our dataset balanced. We removed number of rows where the value of stroke was 0. The balanced dataset consists of 783 patients with stroke and 783 without stroke.

|  |  |  |
| --- | --- | --- |
| **Stroke** | **Imbalanced** | **Balanced** |
| 0 | 42,617 | 800 |
| 1 | 783 | 783 |

Fig. 11 Dataset values after balancing

|  |  |  |  |
| --- | --- | --- | --- |
| **Model used** | **AUC without preprocessing** | **AUC without preprocessing** | **Model improvement with preprocessing** |
| Logistic Regression | 0.84575569 | 0.73613625 | 14.8911% |

Fig. 12 Results for modeling with balanced datasets

# Conclusion

We got an accuracy of 84% which is good, but the model will probably not work well on real world data because we trained our model using very less data when we balanced it. This work on imbalanced dataset shows that machine learning algorithms can be a powerful tool to achieve good results for healthcare problems. Furthermore, we found that the machine learning tools provide a strong mechanism to handle a variety of tasks involving both imputation of missing data and stroke classification.

**References**

1. T. Chen and C. Guestrin, “XGBoost,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, pp. 785–794, 2016.
2. XGBoost “<https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>”
3. <https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data>
4. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>