St Joseph’s COLLEGE (Autonomous), Bengaluru



Heart Stroke Prediction

Advance Statistical Methods Project Report

**Submitted by:**

**Shivanshi-21BDA33**

**Ayesha-21BDA18**

**Aaran-21BDA23**

**Submitted to:**

**Jayati Kaushik**

**Department of Advance Computing**

**St Joseph’s College(Autonomous), Bengaluru**

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** | |
| **TITLE** | **PAGE NO.** |
| **1. INTRODUCTION**  1.1 Introduction and Background  1.2 Definitions  **1.3 Need for study**  **1.4 Objective** |  |
| **2. Dataset**  **2.1 Information regarding dataset**  **2.2 Summary of our dataset**  **2.3 Data cleaning** |  |
| **3. Exploratory Data Analysis** |  |
| **4. Model**  **4.1 Logistic Regression for prediction**  **4.2 Data building and Model training**  **4.3 Anova**  **4.4 Prediction on the testing dataset and Calculate the prediction accuracy** |  |
| **5. Conclusion and Recommendation** |  |
| **REFERENCES:** |  |

**1. Introduction**

**1.1 Introduction and Background**

In today’s world Stroke has been a critical health problem. It badly affects human health and lives. It is the second most deadly disease since 20th century. Stroke is caused as a result of blockage or bleeding of blood vessels which reduces the flow of blood to the brain. Due to this brain does not get sufficient oxygen or nutrients and brain cells starts to die.

Some of the symptoms of stroke are age, overweight, high blood pressure, diabetes, high cholesterol, heart disease etc. In present scenario we come across many people die to heart stroke and we feel that this study may help to some extent to provide the some of the reasons for heart stroke.

**1.2 Definition:**

In 1970, the World Health Organization defined stroke as 'rapidly developed clinical signs of focal (or global) disturbance of cerebral function, lasting more than 24 hours or leading to death, with no apparent cause other than of vascular origin'.

**1.3 Need for study:**

In present scenario we come across many people die to heart stroke and we feel that this study may help to some extent to provide the some of the reasons for heart stroke.

**1.4 Objective:**

As a data analytic students we want to identify the risk factors for stroke. Main objective of our project is to predict whether people will have a stroke and reasons for stroke based on historical data. We are also interested in finding the stroke outcome for potential patients.

* We are trying to find out whether input variables or features affect the stroke outcome.
* We are predicting stroke outcome.

**2. Dataset**

**2.1 Information regarding dataset**

Our dataset showcases person’s body features and their stroke status. It contains 5110 rows with 12 columns. Each row gives the required information about the person. Overall we are finding whether a person is likely to get stroke based on the input parameters such as BMI, hypertension, work type etc.

Below is the attribute Information:

1. id: unique
2. gender: “Male”, “Female” or “Other”
3. age: age of the person
4. hypertension: 0 if the person doesn’t have hypertension, 1 if the person has hypertension
5. heart\_disease: 0 if the patient doesn’t have any heart diseases, 1 if the patient has a heart disease
6. ever\_married: “No” or “Yes”
7. work\_type: “children”, “Govt\_job”, “Never\_worked”, “Private” or “Self-employed”
8. Residence\_type: “Rural” or “Urban”
9. avg\_glucose\_level: average glucose level in blood
10. bmi: body mass index
11. smoking\_status: “formerly smoked”, “never smoked”, “smokes” or “Unknown”\*
12. stroke: 1 if the patient had a stroke or 0 if not

**2.2 Summary of our dataset:**

* There are 9 input variables and 1 outcome(stroke) in the dataset. For our analysis we don’t need ID.
* We can see that some of the columns values are in character, therefore we must change it into factor or number.
* The interesting fact that we must observe is then mean of stroke is 0.04, which means only 4% of the patients have stroke.

**2.3 Data cleaning:**

Firstly we have dropped the column ID.

Secondly we have found that there are many null values in the column BMI. Therefore we have replaced null values with median of the BMI.

In the smoking\_status column we have found that there are large number of unknowns. Therefore we have replaced unknown with most frequent category ‘Never smoked’.

Similarly we examined the columns such as work\_type, residence\_type and ever\_married.

Inorder to perform EDA we have transformed the columns that congains character into binary variables.

**3. Exploratory Data Analysis**

In our dataset we have discrete variables. So we have used barplot to show their distribution.

BARPLOT

From the above barplot we can conclude that female and married people are the majority. Similarly most of them do not have heart disease and hypertension. We can also conclude that private workers and non-smoker are the majority.

For continuous variable we have made use of the histogram.

HISTOGRAM

From the above plots we can say that age is slightly left skewed and glucose level and BMI are right skewed.

One interesting fact we must observe is that the pike in the BMI plot which is the result of replacing null values with median. So we can get an important conclusiom that median is the best estimator than mean as BMI is right skewed.

We have successfully done EDA for individual features. Let us find the association between different variables.I our dataset outcome variable is stroke, so let us find the association between stroke and some of the input features.So some of the plots which we have plotted is boxplot, mosaic plot, heat map.

BOXPLOT

From boxplot we can say that older people are more likely to get stroke. Similarly those who had stroke have higher glucose level and BMI ,but its not that significant.

MOSAIC PLOT

We have used Mosaic plot to find the association between stroke and some discrete input features. Here we can conclude that self-employed workers, those who have hypertension and those who are married are more likely to get stroke. Smoke seems to have little effect on the stroke.

HEATMAP

We have made use of heatmap to knopw the correlation among the inpjut features.Inorder to get heatmap we have converted qualitative variables into quantitative variables.The important factor that we must observe is that the correlation between ever married and age. Here age has highest correlation with stroke.

We have also used scatterplot to find the relationship between features and we have found that people with high BMI and older people are likely to get stroke.

**4. Model**

**4.1 Logistic Regression for prediction**

From heatmap we have seen that there is no certain feature has a strong correlation with stroke. To determine how a certain input feature affect the outcome , we go for a regression. Since stroke is a binary variable we will use logistic regression.

**4.2 Data building and Model training**

As we have considered logistic regression we need to build the training dataset. To test the regression result , we split dataset into training set(70%) and testing set(30%). We have used stroke\_dummy for regression.

From the regression result, we could conclude that:

1. The input features like age, hypertension and work\_type self-employed are statistically significant in the regression results, with p-value smaller than 0.005. Among these age has the lowest p-value.
2. We have noticed in our EDA that age and hypertension are postively correlated with stroke and here we see the confirmation of it.
3. Self-employed is negatively correlated with stroke. So we can say that self-employed would reduce the risk of getting stroke. Probably self-employed people would better enjoy the life and have a healthy lifestyle.

**4.3 Anova**

Now we can run the anova() function on the model to see the deviance of the regression model.

ANOVA shows how features lower the original deviance to residual deviance. Besides three significant features shown in the regression result, avg\_glucose\_level also significantly lowers the deviance. Those features with a large p-value shows that even without these features, the model would explain more or less of the same of total variation.

**4.4 Prediction on the testing dataset and Calculate the prediction accuracy**

To visualize the prediction result, we could use the confusion Matrix function from package caret. We set the prediction result to be 0.5, indicating that if the predicted stroke is larger than 0.5, we believe that this person is likely to get stroke.

From the confusion matrix report, we can see that the accuracy is ------, which is relatively high. If we have a close look we can see that most of the prediction would just predict the outcome to be 0. Therefore we can come to most important conclusion that ‘stroke’ is imbalanced that almost all the outcome to be 0. Therefore it is a difficult task to differentiate those who have stroke and thodse who don’t.

One possible way to get bet result is to change the threshold. Therefore we have plotted a relationship between threshold and accuracy.

We get a result which is not so pleasing to us. Therefore we can conclude that even by raise the accuracy by changing the cutoff threshold. So we have made use of random forest classifier to test the outcome.

Even in the random forest classifier we get the accuracy of ----- and we can see that confusion matrix is similar to logistic regression. Therefore we can say that imbalance of the dataset would greatly affect the prediction result.

So we have used under fitiing method to deal with the imbalanced data.

After under fitting we can see that the result from the confusion matrix of logistic regression shows 75% accuracy and that of random forest shows 77.79%.

We have also made use of the over fitting to deal with the imbalanced data.

The result from the logistic regression shows that the accuracy of 77.79% and random forest classifier shows 99.08% accuracy.

**5. Conclusion and Recommendation**