**ANALYSIS OF STROKE DATA**

INTRODUCTION

The process of cleansing, converting, and modelling data to find relevant information for business decision-making is known as data analysis. Extracting usable information from data and making decisions based on that analysis is the goal of data analysis. We can observe a simple illustration of data analysis whenever we make a decision in our daily lives, evaluating what has happened in the past or what will happen if we take that action. In its simplest form, this process includes looking at the past or future and making a choice based on that analysis.

Stroke happens when there is a pause or a significant reduction in the blood flow to a specific area of the brain. Because of this, brain cells begin to degenerate within minutes because they are unable to receive the oxygen and nutrition they require. According to the World Health Organization (WHO), stroke is the second most common cause of death worldwide. Strokes are a medical emergency that demand immediate care. Life after such a major occurrence can be very difficult for people, even if the chances of achieving a strong recovery are stronger than ever after a stroke. However, there are many things that can be done to improve the quality of life for stroke patients, and in many situations, people are still able to lead active lives.

The platform that I am going to use in this analysis is SAS (Statistical Analysis System). The software program used for data analysis and report authoring is called SAS, or the Statistical Analysis System. SAS is a collection of computer programs that work together to calculate simple and complicated statistical analyses, change data, and create reports. When doing complex analyses, SAS can be used either through the SAS Analyst drop-down menus or by writing one's own code(Rodriguez 2011). We can access raw data files and data from an outside database with easy and almost any data format can be read and written. Use tools for data entry, editing, retrieval, formatting, and conversion to manage data. Also we can utilize descriptive, statistical, multivariate, forecasting, modelling, and linear programming methods to analyze data.

DATA SELECTION AND PREPROCESSING

Health issues are getting worse worldwide these days. It is the area where we must study and increase our preventative measures against these diseases. I obtained the information necessary for this study from the Kaggle platform. The information relates to the chance of stroke conditions. The dataset contains the criteria for the likelihood of having a stroke. It contains the age, gender, hypertension, glucose level, smoking status, married status, work type, heart disease, bmi, residence status. In this data the response variable is stroke and rest are the predictors. During the data preprocessing we evaluate the data carefully and check is there any null values in it. I f there is any null or missing values we need to replace it with any values or need to delete the row or column that contains null values. Since in my dataset there is no null values. So don’t need to do anything with the dataset. Also, I deleted the column bmi which is not necessary for the analysis. The dataset has five categorical columns and four continuous.

PROPOSED SOLUTIONS AND THEIR JUSTIFICATIONS

To make it simple to evaluate a dataset and identify the solutions, we can look at various statistical methods and graphical visualizations while doing an analysis. Here, I utilized the SAS software to check binary logistic regression and the t-test and chi-square with various variables. By using these methods, we can analyze whether a person get a chance of stroke or not. Nowadays, to predict these problems with huge dataset manually is not easy. So, we use these methods to predict the solutions for the problem.

I start by reading the data and looking for any null values. Checking the dataset's normal distribution comes next in the processing. A continuous probability distribution that is symmetrical around its mean is called a normal distribution. For example, in the given figure1. of age data the distribution is at the mean value 43. Similarly, I check this with other continuous columns.

Chart, histogram

Description automatically generatedfigure1

After that, I looked at the summary statistics, which provide the dataset's mean, standard deviation, minimum and maximum values. Here we can see the summary statistics for continuous and categorical values. Figure 2 below shows the dataset's summary statistics. Here we can see one missing value in the column gender which I have replaced with a value. Age has maximum value 82 and minimum value 0.08.

Table

Description automatically generatedfigure2

I evaluated now at frequency by various columns using table analysis. Figure 3 illustrates, for example, that persons who are married experience strokes more frequently than those who are single. The frequency of people who are married is 220 and not married are 29. Women have a higher chance of getting a stroke based on their gender than men do. Smokers and non-smokers both have a chance of suffering a stroke. Regarding employment, those who work in the private sector have a higher probability than others, although those who live in rural areas have a slightly higher chance. Additionally, individuals without heart disease or hypertension have a higher chance of having a stroke.

Table

Description automatically generated with medium confidencefigure3

Correlation analysis between predictors and response variables is the next step. This approach allows us to determine the relationship between one or more variables. Correlation varies from 1 to -1. As we can see, the high degree of correlation is with age compared to the others. The correlation is close to 0.

Text, table

Description automatically generated

Figure4

The relationship between a binary response and a group of predictor variables is modelled using logistic regression. The most popular value is the c value. The concordance statistic, or c, calculates the likelihood that an observation that includes an event will have a higher predicted probability than an observation that does not. The c value is determined by dividing the total number of pairs by the number of concordant results + 1.5 times the number of ties. The possible values fall between 0.5 and 1.0, with 1.0 denoting perfect prediction.

The next method I used is one sample t-test. A statistical hypothesis test called the one-sample t-test is used to examine whether an unknown population mean differs from a given value. In the below figure5, we can see the t value for age is 0.72 and the probability p value as 0.2369 which means our null hypothesis is true and we can accept our null hypothesis. As we know if the p value is greater than significance value 0.05, we accept null hypothesis otherwise we reject it. In this test I give null hypothesis as mu^=43 and sides=u, that is, in the data ages which are greater than 43 have a chance for stroke. Since here the p value is greater than 0.05, so people having age above 43 has the chance for occurring stroke disease. Similarly, when I check t test for glucose level. For this I give null hypothesis as h0=105. People having glucose level greater than 105 has more chance for stroke is my null hypothesis. But my result shows that, the p value of glucose level is 0.0350, that is, p value is less than 0.05. So, we need to reject our null hypothesis. That means people having glucose level less than 105 has more chance for stroke.

Table

Description automatically generatedfigure5Chart, histogram

Description automatically generated

Table

Description automatically generatedfigure6

In figure6, the number 0.611 indicates that average glucose level has a very less capacity for the cause of stroke disease. Similarly, the features like gender, heart disease, smoking, hypertension, ever married and work type has less significance on the response variable. The c value for these criteria is around 0.5. While, in the case of age the c value is 0.835 which is very strong significance for the response variable.

The next thing to look in logistic regression is odds ratio plots. The odds ratio plot graphically presents the outcomes of the odds ratio table. The logistic model's parameter estimates are applied to the predictor values to create this plot, which is then converted to a probability scale. With an odds ratio of 1, the null hypothesis is represented by a reference line. The effect of the variable is not significant when the confidence interval crosses the reference line. In the case of gender, smoking, and residence type the odds ratio is between 0.8 and 1.4 where 1 is included. So, these predictors are significant for the response variable. While, age, heart disease, glucose level, hypertension and ever married crosses the reference line, so, it is not significant. Then we estimate the maximum likelihood for each predictor. The maximum likelihood for gender, smoking and residence, the p value is greater than 0.05 while for age, heart disease, hypertension, glucose, level and ever married the p value is less than 0.05.

Conclusion

Worldwide, stroke continues to be a major contributor to disability and deaths. The burden of stroke has significantly increased globally over the past few decades because of age and population growth as well as a rise in the incidence of lifestyle changes, particularly in low- and middle-income countries. In the coming decades, an increasing number of patients will require the assistance of medical professionals with knowledge of neurological diseases(Abajobir, Abate et al. 2017).

From the analysis of stroke dataset, we can see that age and smoking are the main criteria for causing stroke. Then it comes the gender, residence type, heart disease, hypertension, work type, ever married and glucose level. Through this analysis we are finding out the significant predictor for the response variable and come to a solution for this disease.

Given that smoking is one of the leading causes of stroke, we can use this opportunity to educate people about its use. Social media can be used to spread awareness because it is the fastest informational channel. We can also stage any street performances or awareness campaigns concerning these concerns. Since these criterions are interrelated to each other, by solving any one criterion we can control the rest. Additionally, we may offer patients advice on how to control their blood pressure and glucose levels. Guidelines such as exercise, yoga, and meditation aim to divert people's focus from a significant issue that magnifies these issues.

According to the findings, persons who live in rural areas have a higher risk of stroke. This, in my opinion, is due to the lack of hospitals in this region and the general lack of understanding of stroke disease symptoms. Therefore, we must investigate whether there is a lack of hospital infrastructure in rural areas and raise awareness of the disease. Another contributing reason to this disease is heart disease. To stop further harm, people with heart disease should control what they eat and undergo regular checkups. People with heart problems frequently use blood clotting medicines and stopping these medicines could result in a stroke. They should be conscious not to discontinue taking the medicines that their doctors recommend.

Therefore, people should be aware of the symptoms for any disease and the government should make sure of the awareness about this.

REFERENCES

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

<https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day>

<https://www.guru99.com/what-is-data-analysis.html>

<https://www.medicalert.org.uk/news/2021/08/31/life-following-a-stroke?gclid=CjwKCAjw2OiaBhBSEiwAh2ZSP2HPAhVdanX6LUVMl1nf5dUN8KmB44DDU7efQ7mkWMt5581O2BNICBoCwNgQAvD_BwE>

<http://www2.hawaii.edu/~halina/603/SASintro.pdf>

https://www.guru99.com/sas-tutorial.html#2

Abajobir, A., et al. (2017). "Neurological Disorders Collaborator Group, Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015." Lancet Neurol **16**: 877-897.

Rodriguez, R. N. (2011). "Sas." Wiley Interdisciplinary Reviews: Computational Statistics **3**(1): 1-11.

Appendix

Uploading data and reading it

A picture containing text

Description automatically generated

Dropping a column

Graphical user interface, text, application

Description automatically generated

Summary Statistics

Text

Description automatically generated

Distribution Analysis

Text, letter

Description automatically generated

Table Analysis

Graphical user interface, text, application, email

Description automatically generated

Correlation Analysis

Graphical user interface, text, application

Description automatically generated

T-test

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Binary Logistic Regression

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated