PROJECT REPORT

DEVELOPING A PREDICTION MODEL FOR IDENTIFYING HIGH-RISK INDIVIDUALS FOR STROKE USING DEMOGRAPHIC AND HEALTH DATA

STATS STUDENTS

Kranthi Bathini, Venkata Siva Rao Kamisetty, Joshua Gaddam, Varshini Singamaneni, Naga Shalini Tanneru

Introduction: A brain stroke is a critical medical condition resulting from an interruption in the blood flow to the brain, leading to potential brain damage and long-term negative impacts on the patient's health. stroke is the second most common cause of death. (Kuriakose & Xiao, 2020, p. 2). Given that it is a significant contributor to death and disability globally, early detection and treatment are essential to improving patient outcomes. Our goal is to create a prediction model based on available health data to effectively and accurately identify individuals who are at high risk of experiencing a stroke. According to Healthline (2021), A brain stroke is a serious medical situation that happens when the blood flow to the brain is hindered or decreased, resulting in a lack of oxygen and essential nutrients to the brain cells. This lack of oxygen and nutrients can cause the brain cells to die in just a few minutes. (Brain Stroke: Symptoms, Causes, Treatment, and Prevention, n.d.).

Problem Statement: The purpose of this study is to assess the risk factors for brain stroke in individuals and determine whether there are any appreciable variations in risk among various demographic groups. The project also aims to identify the most prevalent medical disorders or comorbidities that co-occur with stroke in the dataset and investigate whether there are any significant connections between these diseases and stroke.

Research questions:

We specifically want to investigate into the following questions:

1. Is there a correlation between age and the likelihood of stroke? Does the risk of stroke increase as a person gets older, and if yes, at what rate?

2. Do men and women have different chances of experiencing a stroke?

3. Are married individuals more likely to suffer from stroke than unmarried individuals?

4. Does smoking status have an impact on the risk of stroke?

Explaining the Dataset:

Link for dataset: https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset

The Brain Stroke Dataset is a comprehensive collection of medical records and imaging data from stroke patients. The dataset includes 4981 observations (rows) and 11 variables. Including demographic information such as age, gender, and smoking status, and clinical features such as hypertension, heart disease, and glucose levels.

The dataset contains a variable that indicates whether a patient has experienced a stroke or not. The variable "stroke" is binary, with a value of 1 indicating that a patient has had a stroke, and a value of 0 indicating that they have not.

The variables in dataset are:

The Brain Stroke Dataset is a comprehensive collection of medical records and imaging data from stroke patients. The dataset includes 4981 observations (rows) and 11 variables.

1. gender: The gender of the patient (Male, Female, or Other).
2. age: The age of the patient in years.
3. hypertension: A binary variable indicating whether the patient has hypertension (1) or not (0).
4. heart_disease: A binary variable indicating whether the patient has heart disease (1) or not (0).
5. ever_married: A binary variable indicating whether the patient has ever been married yes or no.
6. work_type: The type of work the patient does (Private, Self-employed, Govt_job, children, or Never_worked).
7. Residence_type: The type of residence of the patient (Urban or Rural).
8. avg_glucose_level: A numerical variable indicating the average glucose level in the patient's blood.
9. bmi: A numerical variable indicating the body mass index (BMI) of the patient.
10. smoking_status: A categorical variable indicating the smoking status of the patient (formerly smoked, never smoked, or smokes).
11. stroke: A binary variable indicating whether the patient had a stroke (1) or not (0).

Feature selection:

Descriptive Statistics:

- The minimum age is 0.08, the maximum age is 82, and the median age is 45. The mean age is 43.42.
- There are 249 people who had a stroke (the minimum value is 0 and the maximum value is 1). The mean value for stroke is 0.04979, indicating that strokes are rare in this dataset.

Data Visualization:

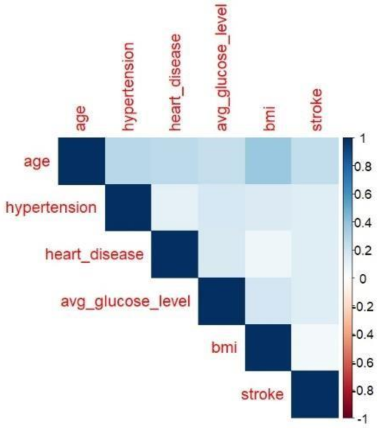
Histogram: We can forecast the stroke status by age using the histogram. According to histogram, the age group between 30 to >80 have more probability of stroke.

Pie chart: we generated pie charts to visualize the distribution of each categorical variable in the data frame df_cat. For each variable in the data frame, the code creates a table of counts for each category, calculates the percentage of each category, and creates a pie chart using the ggplot2 package.

Correlation analysis:

We coded for a correlation matrix of the numeric variables in the data frame df, that extracts the correlations between the variable "stroke" and the other numeric variables and sorts them in descending order. The **corrplot** function from the **corrplot** library is used to visualize the correlation matrix with a color scheme. The output shows the correlations between the variables, with stronger correlations displayed in brighter colours.

The highest correlation is between age and stroke, with a correlation coefficient of 0.246, indicating a moderately positive correlation. This means that we can predict as age increases, the likelihood of having a stroke also increases. The variables heart disease, avg_glucose_level, and hypertension also show moderate positive correlations with stroke, with correlation coefficients of 0.135, 0.133, and 0.132, respectively.

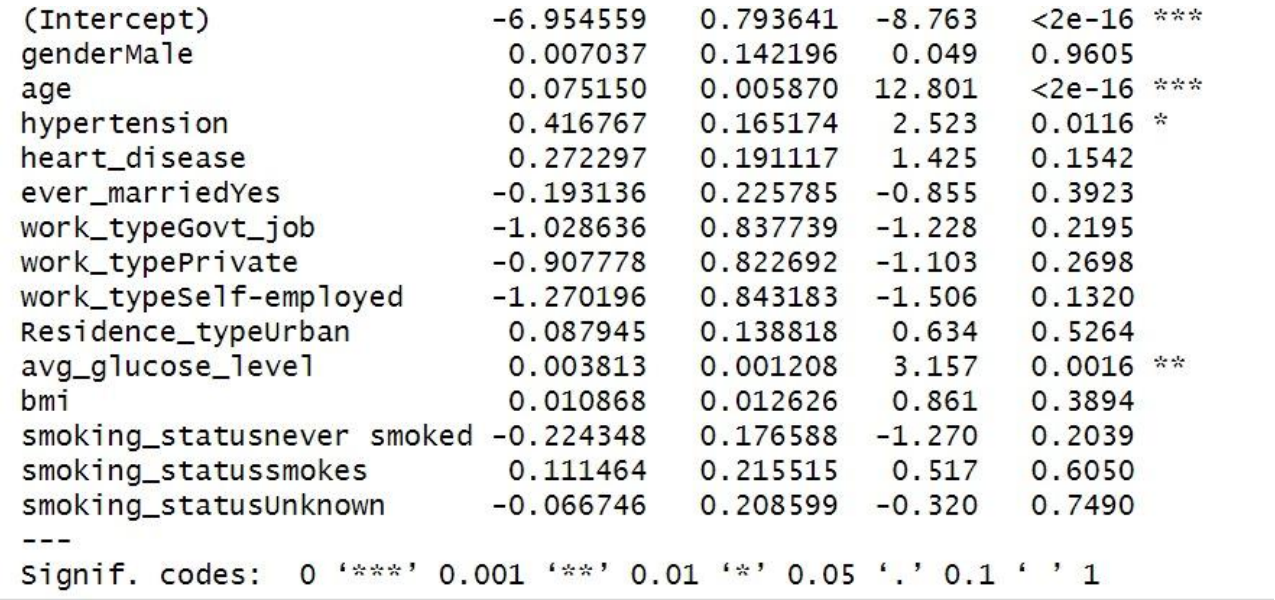


Model Development: Logistic Regression

We performed a logistic regression model to predict stroke occurrence based on various predictors such as gender, age, hypertension, heart disease, ever married status, work type, residence type, average glucose level, BMI, and smoking status. We then printed the summary of the model and extracted the names of significant predictor variables at a significance level of

0.05. This information can be used to understand which variables are most strongly associated with the occurrence of stroke.

The results indicated:



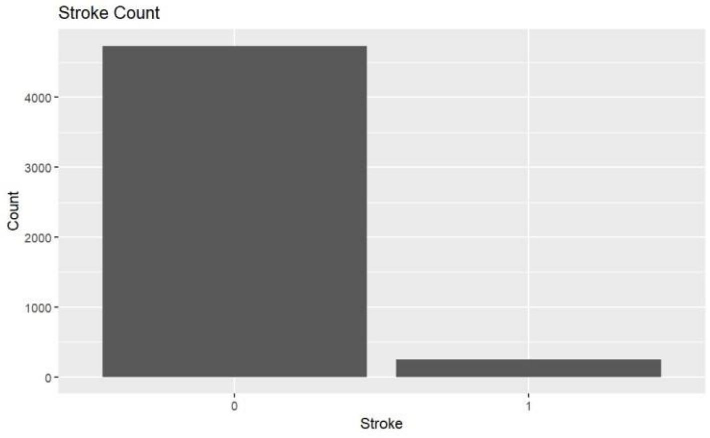
```
(Intercept)               -6.954559   0.793641  -8.763   <2e-16 ***
genderMale                 0.007037   0.142196   0.049   0.9605
age                        0.075150   0.005870  12.801   <2e-16 ***
hypertension               0.416767   0.165174   2.523   0.0116 *
heart_disease              0.272297   0.191117   1.425   0.1542
ever_marriedYes           -0.193136   0.225785  -0.855   0.3923
work_typeGovt_job         -1.028636   0.837739  -1.228   0.2195
work_typePrivate          -0.907778   0.822692  -1.103   0.2698
work_typeSelf-employed    -1.270196   0.843183  -1.506   0.1320
Residence_typeUrban        0.087945   0.138818   0.634   0.5264
avg_glucose_level          0.003813   0.001208   3.157   0.0016 **
bmi                        0.010868   0.012626   0.861   0.3894
smoking_statusnever smoked -0.224348   0.176588  -1.270   0.2039
smoking_statussmokes       0.111464   0.215515   0.517   0.6050
smoking_statusUnknown     -0.066746   0.208599  -0.320   0.7490
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The logistic regression model shows that **age, hypertension, and average glucose level** have a significant positive association with the likelihood of stroke, with p-values less than 0.05.

Prediction models:

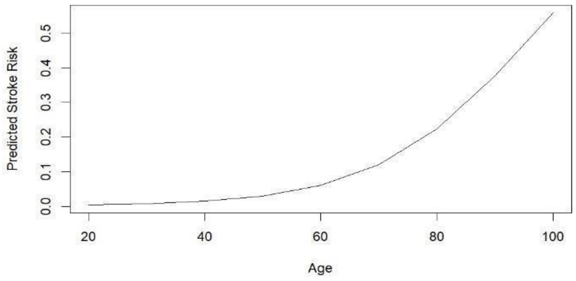
Number of people in the data had a stroke before:

We created a bar plot of the stroke variable in the df data frame. The sum function is then used to calculate the number of strokes that did not occur (stroke=0) and the number of strokes that occurred (stroke=1). This can be useful for understanding the distribution of the outcome variable and identifying any imbalances in the data set.



Estimating the stroke risk at different ages:

This plot can be used to visualize the relationship between age and stroke risk and to identify age groups with higher predicted risk of stroke. the plot shows the relationship between age and predicted stroke risk using a line plot. The x-axis shows the age, and the y-axis shows the predicted stroke risk.



From the plot we can predict that with the increase in age the probability of stroke is increased.

RQs:

Is there a correlation between age and the likelihood of stroke? Does the risk of stroke increase as a person gets older, and if yes, at what rate?

Null hypothesis: There is no relationship between age and the risk of stroke. Any observed association is due to chance.

Alternative hypothesis: There is a positive relationship between age and the risk of stroke. As age increases, the risk of stroke also increases.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.204396   0.336974  -21.38   <2e-16 *** age
0.074462   0.004946   15.06   <2e-16 ***
```

The "Pr(>|z|)" column shows the p-value for the coefficient, in this case, both coefficients are highly significant (p < 0.001), thus, we reject null hypothesis indicating a strong association between age and the likelihood of having a stroke.

By taking the exponent of the estimated standard deviation of age and subtracting 1 from it, we calculated the percentage increase in the probability of having a stroke with each one-unit increase in age, we found that the rate of increase is 0.07730.

Chi square test to predict gender vs stroke:

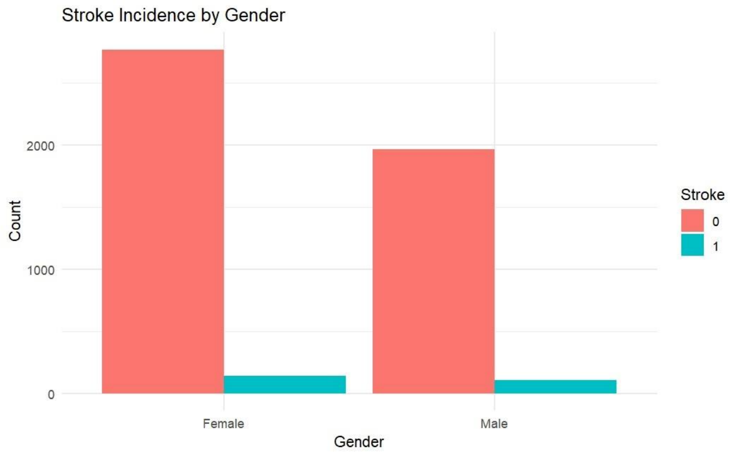
Null hypothesis: There is no difference in stroke risk between men and women.

Alternative hypothesis: There is a difference in stroke risk between men and women.

```
data:  cont_table
X-squared = 0.31352, df = 1, p-value = 0.5755
```

Since **P value > 0.05**

So, we **fail to reject the null hypothesis** therefore, we can there is no significant relationship between gender and stroke



Men and women have equal probability.

Chi square test to predict ever married vs stroke:

Null hypothesis: There is no significant difference in stroke risk between married and unmarried individuals.

Alternative hypothesis: Married individuals have a higher (or lower) risk of stroke compared to unmarried individuals.

```
Pearson's Chi-squared test with Yates' continuity correction

data:  table
X-squared = 57.481, df = 1, p-value = 3.412e-14
```

Since **P < 0.05** So, we **reject** the null hypothesis and conclude that there is a significant relation between marital status and stroke.

Chi square test to predict smoking vs stroke:

Null hypothesis: Smoking status has no impact on the risk of stroke.

Alternative hypothesis: Smoking status has an impact on the risk of stroke.

```
Pearson's Chi-squared test

data:  smoking_table
X-squared = 28.734, df = 3, p-value = 2.548e-06
```

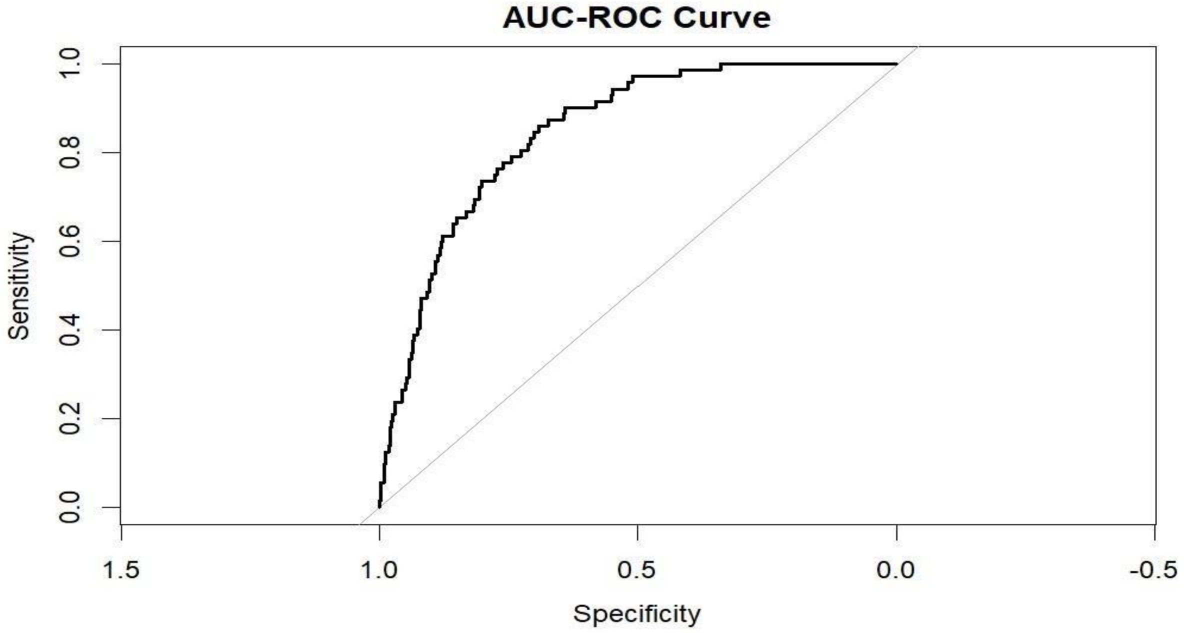
Since, **P < 0.05**, So, we can reject null hypothesis and conclude that smoking has impact on stroke i.e., there is a relation between smoking and stroke.

Model building and Model validation:

In this analysis, we used logistic regression to predict the occurrence of stroke based on various factors such as "gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status". Logistic regression is a statistical method used to calculate odds ratio when there are multiple independent variables. (Sperandei, 2014 p. 12). The data was split into training and testing sets, and the model was trained on the training set. The accuracy of the model on the testing set was found to be **0.95**. The precision and recall of the model were Nan and 0, respectively, due to lack of more positive values in the dataset. This indicates that the model is good at identifying true positive cases. The area under the ROC curve (AUC-ROC) was found to be **0.85**, indicating that the model performs well in distinguishing between positive and negative cases. Overall, the logistic regression model shows promise for predicting stroke occurrence based on the given set of variables.

```
Accuracy: 0.95
Precision: NaN
Recall: 0
F1 Score: NaN
AUC-ROC: 0.85
```

AUC-ROC CURVE:



CONCLUSION:

This project aimed to explore the associations between age, gender, marital status, smoking, and stroke risk. The results revealed that age is positively correlated with stroke risk, while no significant difference was found between men and women or married and unmarried individuals. However, smoking was found to be a significant risk factor for stroke, indicating that efforts to reduce smoking prevalence could potentially reduce the incidence of stroke. These findings underscore the need for targeted public health interventions, particularly in older age groups, to mitigate the risk of stroke and improve overall health outcomes.

Limitations:

The dataset contains only 4981 observations, which may not be representative of the general population or certain subgroups. This could limit the generalizability of any findings or conclusions drawn from the data. The dataset includes information on whether or not a patient had a stroke but does not include detailed information on the severity or type of stroke, which could limit the ability to draw conclusions about specific aspects of stroke risk or prognosis.

References:

Healthline. (2021). Brain Stroke: Symptoms, Causes, Treatment, and Prevention. Retrieved from https://www.healthline.com/health/stroke

Kuriakose, D., & Xiao, Z. (2020). Pathophysiology and treatment of stroke: Present status and future perspectives. *International journal of molecular sciences*, *21*(20), 7609. https://doi.org/10.3390/ijms21207609

Sperandei S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12–18. https://doi.org/10.11613/BM.2014.003