FIT5145

# Introduction to Data Science Assignment – 3

Stroke Risk Prediction for Healthcare

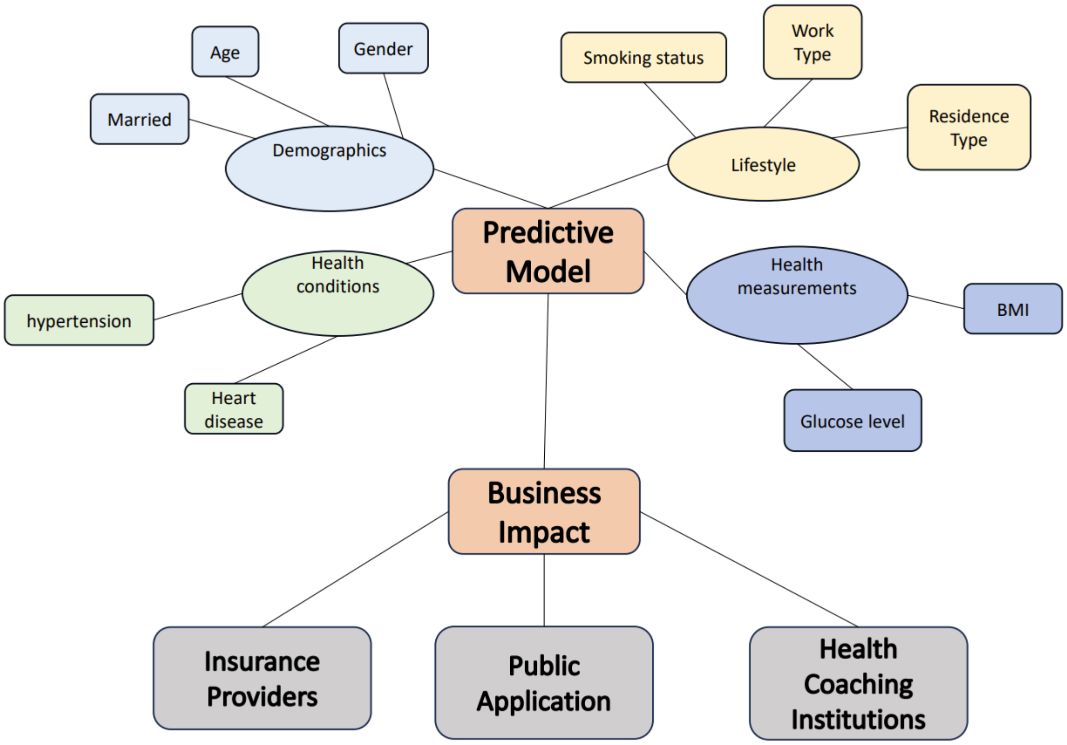
Jeet Faldu
STUDENT ID: 33670021

# Contents

- ## Project Description:

    - In this case study, we will explore a data science project which will use stroke prediction dataset to predict people's chance of having a stroke based on their lifestyle and demographics. This prediction will be helpful for Insurance providers in identifying such profiles before approval. Additionally, it will benefit the public as they can monitor their chance of having a stroke beforehand, which can help them to take preventive measures.

    - Different machine learning models will be used on the public's demographic, lifestyle and various human health conditions and measurements. Considering those parameters our model will predict if a given applicant can have a stroke or not.



*Business Model*

    - ### Data Roles:
        - Data Scientist: A data scientist will be responsible for data extracting, transforming, and loading data. Additionally, perform data exploration, feature engineering and model selection and validation. Furthermore, building a predictive model based on machine learning.
        - Business Analyst: A business analyst will be responsible for understanding the business needs and deriving them from the predictive model.
        - IT Support: Provides the assistance needed for the deployment of predictive models on client's system and assures smooth integration in respective environments.

- Business Case:
    - Project Aim:
        - As provided by National Library of Medicine “Stroke is the leading cause of long-term adult disability and the fifth leading cause of death in the US, with approximately 795,000 stroke events in the US each year” [1]. Stroke is a huge concern for society as it can lead to human loss. Additionally, it can cause long-term impairment. Stroke is serious problem globally as it is fourth leading cause of death worldwide.
    - Business impact:
        - Insurance Providers:
            - Insurance providers can use this model to identify the new applicant and take decisions based on this model that particular customer will be given a high coverage plan or not.
        - Public Application:
            - Common people can get benefit of this model by integrating these models with their daily fitness application or health application. People can get notified with this integration if they fall under the risk of stroke due to changes in their lifestyle or any features which are considered while building this model. This gives an advantage to users which can help them to take preventive measures and aim to change their lifestyle to reduce risk of Stroke.
        - Health Coaching Institutions:
            - Health Coaching Institutions can use these models to identify the target customers which are at risk of developing characteristics which can lead to Stroke in feature.
    - Challenges:
        - If real time data is included than huge volume of data from multiple sources will be generated
        - Different parameters might flow through real time data and feature selection and exploratory analysis needs to be done again to check if relevant dependencies of stroke are present on new variables.
        - Consistency of the models need to be maintained as false positive or false negative results can directly impact patients’ life.

## Characterising and Analysing Data:

### Data Sources:

Electronic Health Records (EHR): EHR obtained from hospitals and clinics contains patient's medical conditions including medical history and lab results.

Medical Imaging: Important information regarding the condition of blood vessels can be obtained from medical imaging data such as MRI and CT scans.

Patient Surveys: Surveys conducted at multiple sources can be helpful in collecting patient's demographic information like lifestyle, habits, and family history.

Wearable Devices: Devices like smart watches can provide patient's routine data like physical activities, heart rate, sleep cycle etc.

Online Data Repositories : Kaggle

## Volume:

Volume is the amount of data that is considered at the initial phase of the analysis. Size and amount vary with time. Dataset contains information about patients whether they have stroke condition or not. Dataset contains ~5k records and 11 columns.

## Velocity:

Velocity of data is how often data is generated at source and it is sent to consumers which utilize this data for their analysis. For e.g., Healthcare data generated of present time reaches to organizations after processing in 2 weeks than velocity of dataset can be considered as 15 days. For this project I have considered static data sources.

## Variety:

Variety refers to different data types of data. The dataset considered for this project is tabular data and it contains numerical and categorical information about patient demographics and lifestyle.

## Veracity:

Quality and accuracy of the data can be considered as the veracity of the data. In the current dataset BMI contains few N/A values (~200 records). However, other columns contain non null values and provide information based on the nature of the column.

## Proposed data analysis:

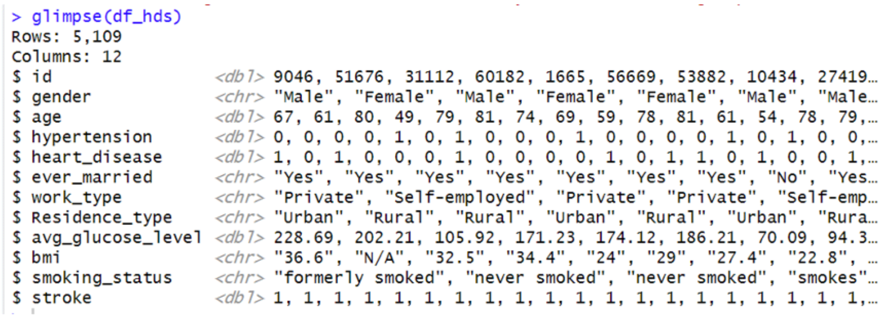
- Exploratory data analysis
  - Data checking and wrangling:
    - Loading the dataset and checking if there are missing values.
    - Checking datatypes and data consistency of each column
  - Data Exploration
    - Numerical variables distribution and visualization
    - Analyzing the relation between Stroke and other Categorical Variables. Visualization of relationships
- Predictive Modeling
  - Decision Tree
    - Implementation of decision tree to predict if new patient can have stroke or not.
  - Logistic Regression
    - Fitting model and predicting on test data for target variable stroke.

# Demonstration

## Exploratory data analysis

### Data checking and Wrangling:

- Considering the nature of healthcare data bigdata tools like Pyspark, Alteryx or AWS, Hadoop for real time data are required for the ETL as multiple data files from multiple resources needs to be handled to perform analysis. The current dataset consists of ~5k records and one data file is present, and analysis have been performed on R studio.
- R has been used for entire analysis and visualization in this project. For data checking, data set is handled in form of data frames.
- To get basic understanding of the column's 'glimpse' method was used to check the data types and few values of the column.



```
> glimpse(df_hds)
Rows: 5,109
Columns: 12
$ id                <dbl> 9046, 51676, 31112, 60182, 1665, 56669, 53882, 10434, 27419…
$ gender            <chr> "Male", "Female", "Male", "Female", "Female", "Male", "Male…
$ age               <dbl> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, 78, 79,…
$ hypertension      <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0,…
$ heart_disease     <dbl> 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1,…
$ ever_married      <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", "Yes…
$ work_type         <chr> "Private", "Self-employed", "Private", "Private", "Self-emp…
$ Residence_type    <chr> "Urban", "Rural", "Rural", "Urban", "Rural", "Urban", "Rura…
$ avg_glucose_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.09, 94.3…
$ bmi               <chr> "36.6", "N/A", "32.5", "34.4", "24", "29", "27.4", "22.8", …
$ smoking_status    <chr> "formerly smoked", "never smoked", "never smoked", "smokes"…
$ stroke            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
```

*Figure 1 Column glimpse*

From Figure 1 it is observed that columns have numeric and string datatypes. For all the columns except BMI datatypes make sense but for BMI it should be numeric, but it is flowing as string due to string 'NA' values present in the data. To overcome this issue string 'NA's were converted as Null values and using 'vis_dat' library plotted the data frame to check if there are any null values in other columns.

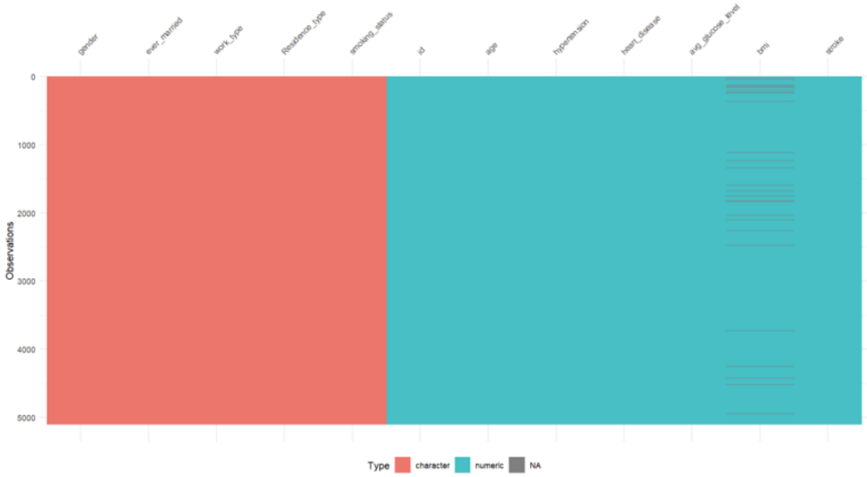


*Figure 2 Null value check*

From Figure 2 it can be observed that apart from BMI column Null values are not present in any columns in the data frame. Further observed that only few records (~200) have null values in BMI column so dropped those records with null values from the data frame.

## Data Exploration

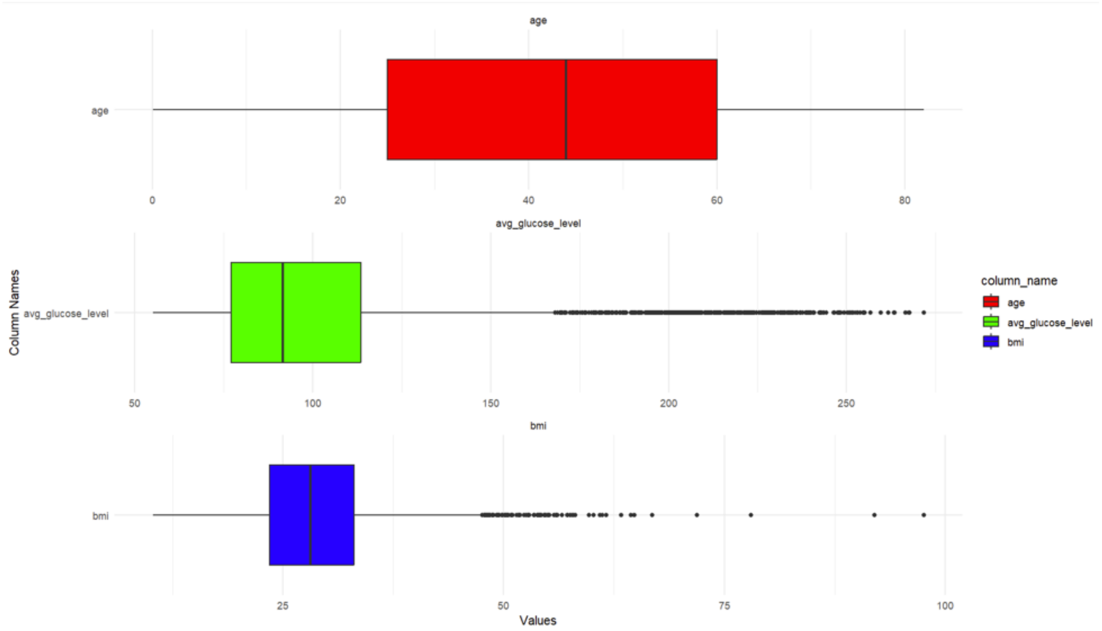


*Figure 3 Box plot for numerical variables*

Boxplots in Figure 3 represent the distribution of the numerical columns in the dataset. Boxplot for age column represents almost symmetric distribution. However, for Average Glucose level and BMI it is observed that distribution is right skewed. However, we can't exclude these outliers because these might be patients who can be identified as stroke patients.

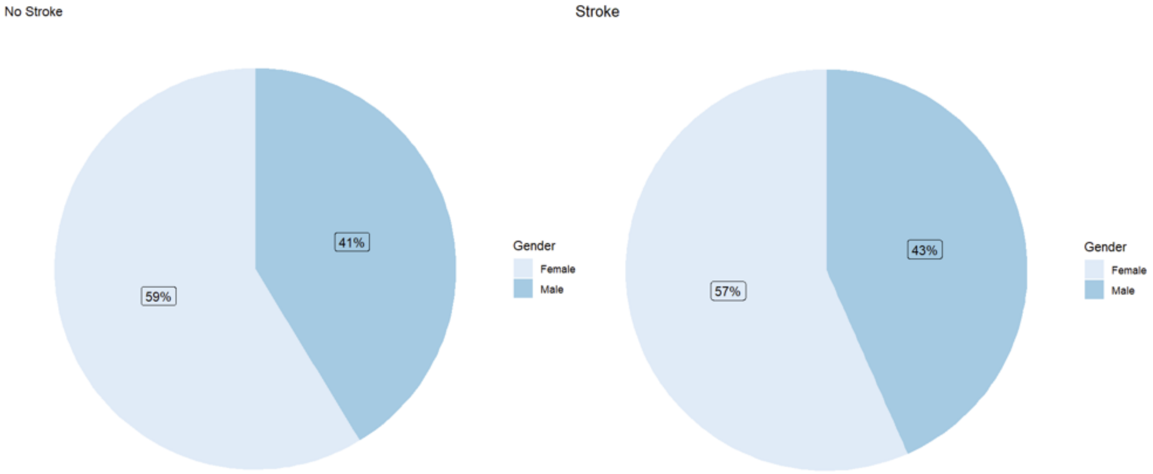


*Figure 4 Gender Distribution*

Figure 4 represents the patients having stroke and their gender. It is observed from the Figure 4 that distribution of male and female patients with and without stroke is almost similar as there is no significance difference in both the plots.

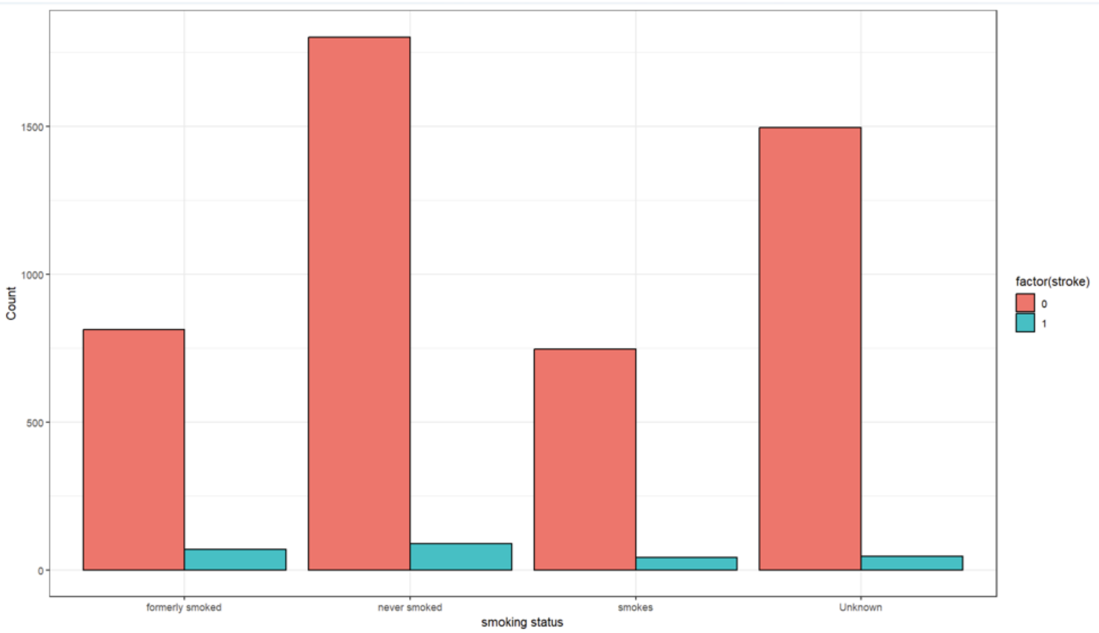


*Figure 5 Smoking Status*

Figure 5 shows the distribution of patients with Stroke flag 1 and 0 based on their smoking habits and it is observed that there is some difference in number of patients with stroke who never smoked or formerly smoked with patients whose smoking status is unknown or ones who smokes.

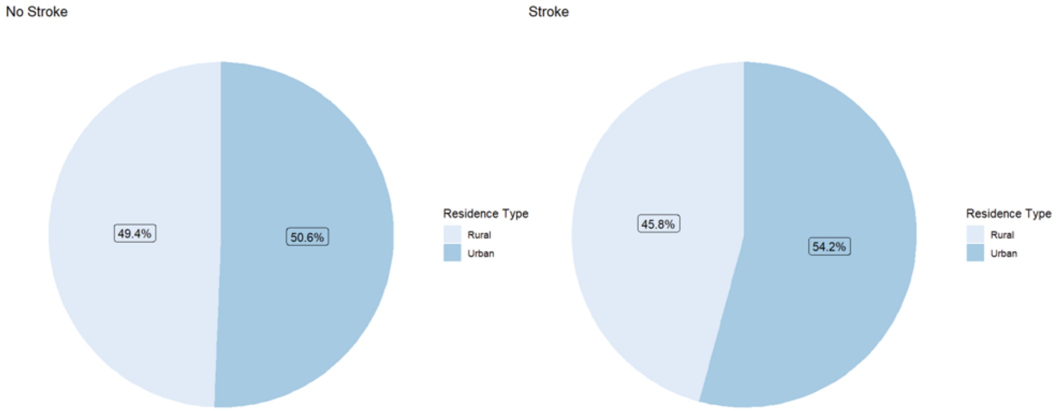


*Figure 6 Residence Type*

Stroke patients based on their residence type is shown in the Figure 6 where it can be observed that there is little difference between patients having stroke i.e., number of patients living in urban area are more compared to number of patients living in rural areas.

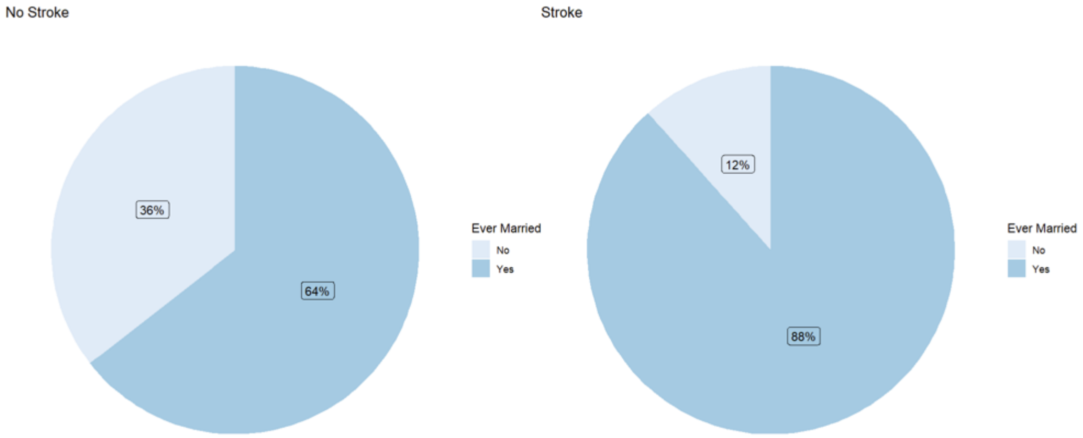


*Figure 7 Marital status*

Figure 7 shows the difference between patients having stroke or not based on their marital status and it is observed that there are a greater number of patients with stroke and are married in comparison with patients having stroke but are not married.

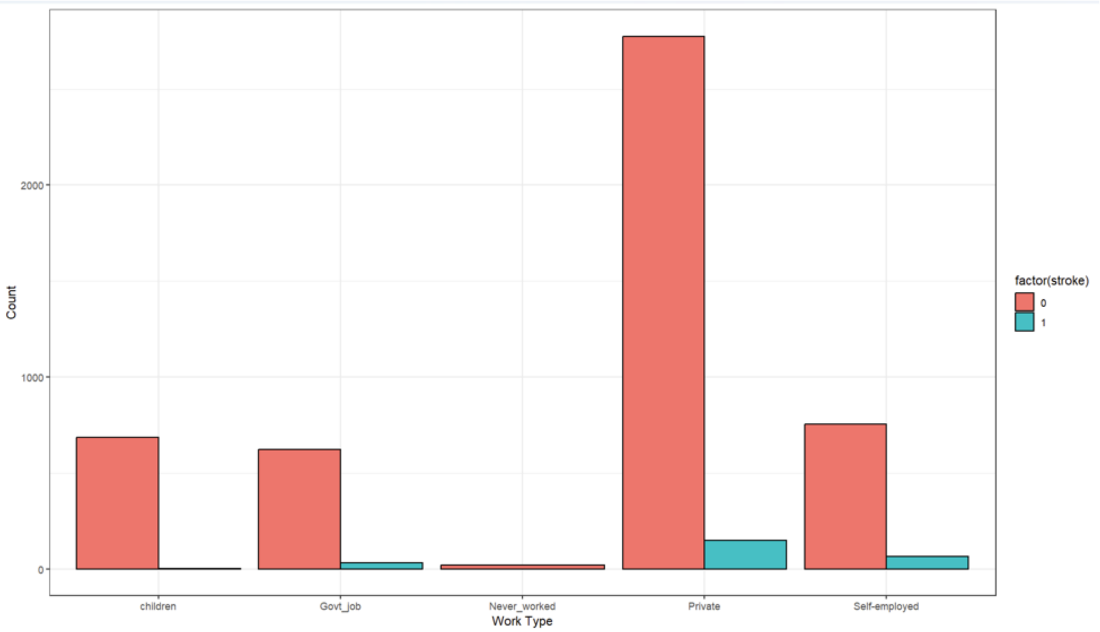


*Figure 8 Work type.*

Patients having Stroke or not based on their work type is shown in figure 8 and it is observed that if we exclude children and patients who have never worked than patients having Private job have a greater number of stroke patients followed by self-employed and Government job.

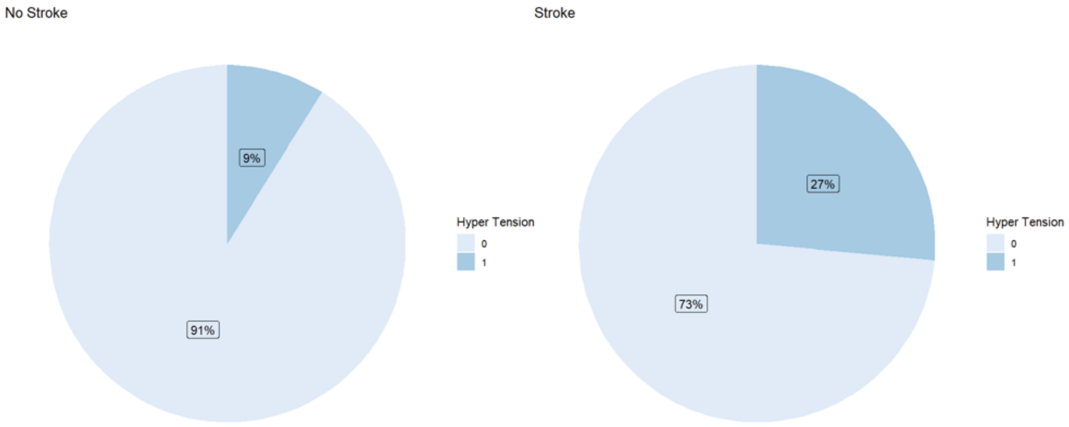


*Figure 9 Hypertension*

Figure 9 shows patients having stroke or not and if they are identified as hypertension patients or not. It can be observed from Figure 9 that among the patients having stroke a greater number of patients are identified as Hypertension patients. This factor can be considered for further analysis in predicting stroke for new patients.

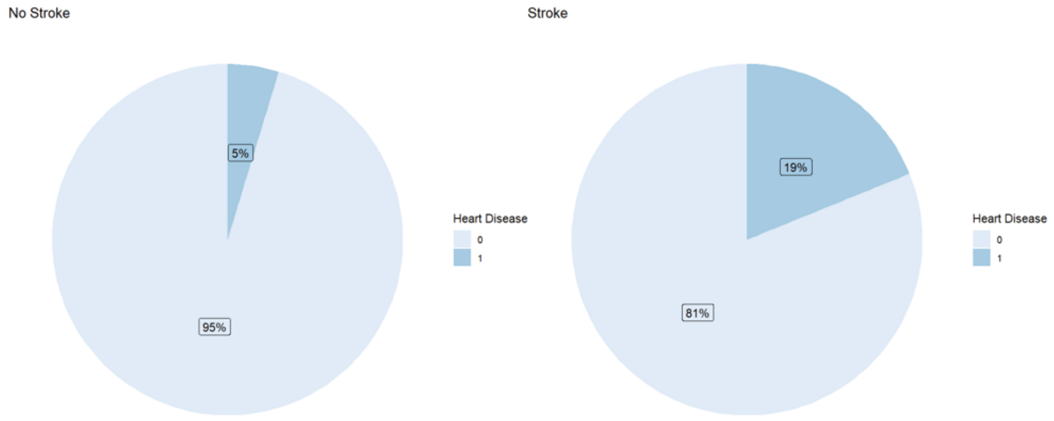


*Figure 10 heart disease*

Patients having stroke or not and the ones who have heart disease or not are shown in Figure 10. From figure 10 it is observed that patients having heart diseases are more likely to have stroke as it can be seen from the right chart that among patients having stroke 19% have heart diseases.

## Predictive Modeling

Initially checked the correlation of Stroke with all the parameters but it was observed that there is no column which is significantly correlated with variable stroke.

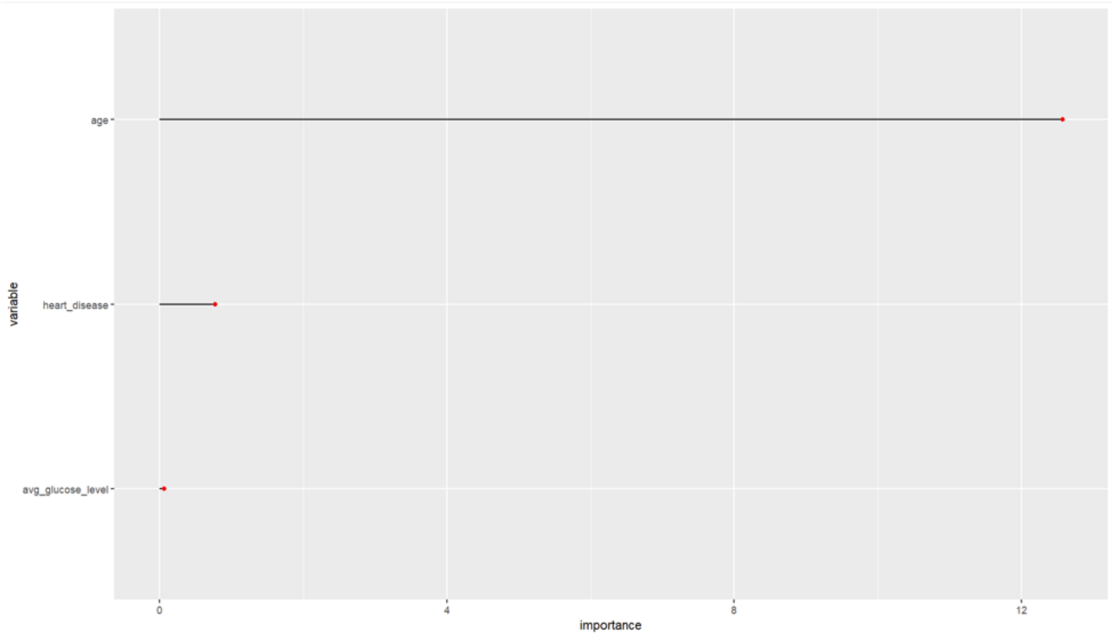


*Figure 11 Importance matrix*

Further to check the important variable for stroke, created plot for importance matrix which is shown in figure 11 and observed that age, heart disease and average glucose levels show importance. Additionally, it is observed that the importance of age variable is significantly higher than heart disease and Average glucose levels.

## Decision Tree:

Initial plan was to use Decision tree by training the model based on the available train dataset and use the model to check accuracy based on the test dataset and further use the model to predict for new patient whether they could have stroke or not.

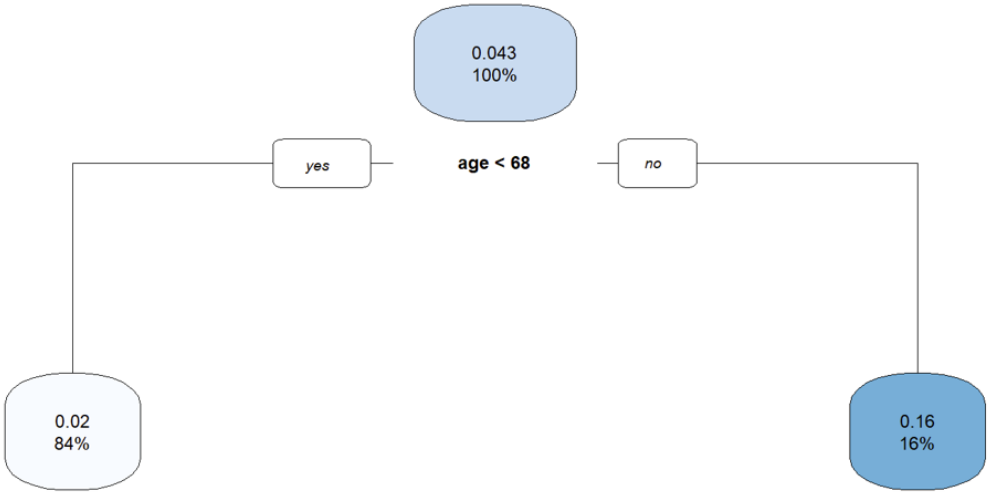


*Figure 12 Decision Tree*

However, it can be observed from figure 12 that the decision tree is created with one root node based on the age variable. This might be the case where purity of the tree was achieved at the root node and the tree which was expected to consider other variables were not included in the tree.

## Logistic Regression:

For the binomial target variable logistic regression is a better algorithm as it assigns probability to an event occurring belonging to certain range. On analyzing the summary of model, it was observed that variables age, Average glucose level, Hypertension and Heart disease have P values closer to 0 thus we will reject the null Hypothesis. Hence, we will consider these variables for predictive modeling.

To check the accuracy of the model threshold of 0.5 was considered. Actual stroke value was compared with the predicted values, and it was observed that model has 95% accuracy.

## Standard for Data Science Process, Data Governance & Management

### Data Science Process Standard

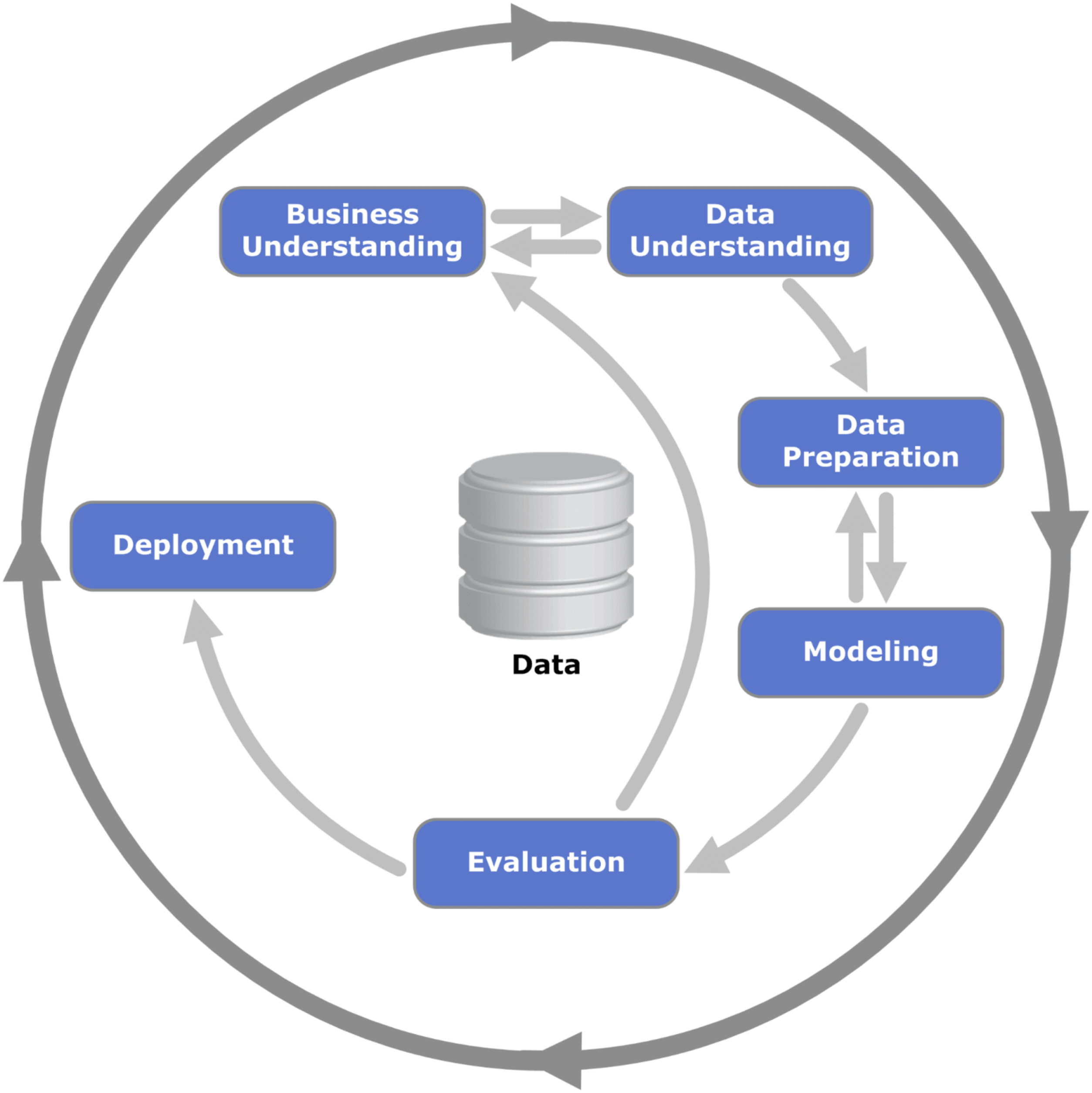


*Figure 13*

The project follows different stages as shown in the Figure 13:

- Business Understanding
  - Main objective of identifying the patients if they can have stroke has been defined.
- Data Understanding
  - Stroke data has been obtained from secondary sources and exploration has been done on the collected data.

- Data Preparation
    - Data checking, data wrangling and exploratory data analysis have been performed.
- Modelling
    - Decision tree and Logistic Regression have been used to predict if patient can have stroke.
- Evaluation
    - Evaluated predicted model and obtained high accuracy.
- Deployment
    - Proposing model for user applications from different domains.

## Data Governance and Management

### Accessibility

- Maintaining an extensive data catalogue that includes data sources, types, and definitions and is easily accessible to important stakeholders is necessary for accessibility.

### Security

- Sensitive health information must be protected from unauthorized access, which requires strict security measures including encryption, access control, and regular audits.

### Confidentiality

- Due to the sensitive nature of the material, confidentiality is of the utmost importance, and personal data is protected through techniques of masking.

### Data Retention

- By routinely evaluating and archiving data, a well-defined data retention policy must be designed to reduce overhead costs and guarantee compliance.

### Ethical Concerns

- Transparency, explicability, informed permission, and adherence to ethical and legal requirements, such as data protection legislation, are all required by ethical considerations.

References:

1) FEDESORIANO. (2021). *Stroke Prediction Dataset*. Www.kaggle.com. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

2) Boehme, A. K., Esenwa, C., & Elkind, M. S. V. (2017). Stroke Risk Factors, Genetics, and Prevention. *Circulation Research*, *120*(3), 472–495. https://doi.org/10.1161/CIRCRESAHA.116.308398

3) Gillis, A. (2021, March). *The 5 V's of Big Data*. SearchDataManagement. https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data

4) Zach. (2022, April 12). *How to Split Data into Training & Test Sets in R (3 Methods)*. Statology. https://www.statology.org/train-test-split-r/

5) Howell, E. (2023, August 27). *The Essence of Logistic Regression*. Medium. https://towardsdatascience.com/the-essence-of-logistic-regression-e9188625cb7d

6) *The Data Science Life Cycle: A New Standard for Operationalizing Data Science*. (n.d.). KNIME. Retrieved October 17, 2023, from https://www.knime.com/blog/the-data-science-life-cycle-a-new-standard