EDA Final Project Report
Topic: Stroke Prediction Analysis

Teammates:

Meghana Boinpally
Raghav Chegu Shyam Kumar
Shefali Mahendra Luley

Dataset:

The dataset being used for the Final Project is the Stroke Prediction Dataset available on Kaggle. This dataset is used to predict whether a patient is likely to get a stroke based on various parameters described below:

Categorical:

- **gender:** "Male", "Female" or "Other"
- **ever_married:** "No" or "Yes"
- **work_type:** "children", "Govt_job", "Never_worked", "Private" or "Self-employed"
- **Residence_type:** "Rural" or "Urban"
- **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"
- **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease.

Numerical:

- **age:** age of the patient
- **avg_glucose_level:** average glucose level in blood
- **bmi:** body mass index

Target Variable:

- **stroke:**Output [1: Patient had a stroke ,0: Patient didn't have a stroke]

Research Questions:

Our project's aim is to visualize and comprehend the risk factors of stroke.

1) **What are the attributes that are associated with stroke?**
2) **Can we perform Dimensionality reduction of this high dimension data by minimizing information loss?**
3) **Will non-linear classifiers be more effective in predicting if an individual has a stroke or not?**

Data Exploration and Cleaning:

On exploration of the data, we found that the data is highly imbalanced, and some inconsistencies shown below:

- There are **4908** observations, **209** with stroke and **4699** without stroke.
- There are **201** records in the BMI column consisting of "N/A".
- The gender column consisting of just one record showcasing the "Other" gender.

As all these inconsistencies accounted for around **4%** of the entire data, we removed them for better analysis. Moreover, for majority of the analysis, we have proceeded to approach the different analysis using the stroke and non-stroke cohort separately, thus, giving a better outlook on the distribution. We also used percentages in the distributions to observe the data better.

Univariate Analysis

Continuous Variables:

As our dataset is heavily imbalanced, we divided our dataset into 2 cohorts.
Cohort 0 having stroke = 0 and cohort 1 having stroke = 1. The below three graphs show the distributions of Average Glucose level, Age and BMI respectively for stroke and non-stroke members. The following observations can be made:

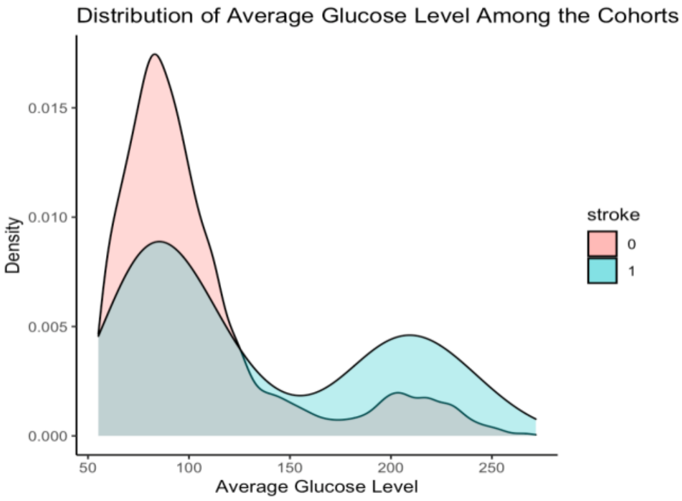


Figure 1

Average Glucose Level

- Figure 1 shows us the distribution of Average Glucose Level across different cohorts
- The distributions of the stroke and non-stroke members are both bimodal
- The difference can be seen in the ranges of 55-110 and 150-280.
- The non-stroke members have relatively lower average glucose level than the stroke members suggesting that it could be a good factor in predicting stroke.

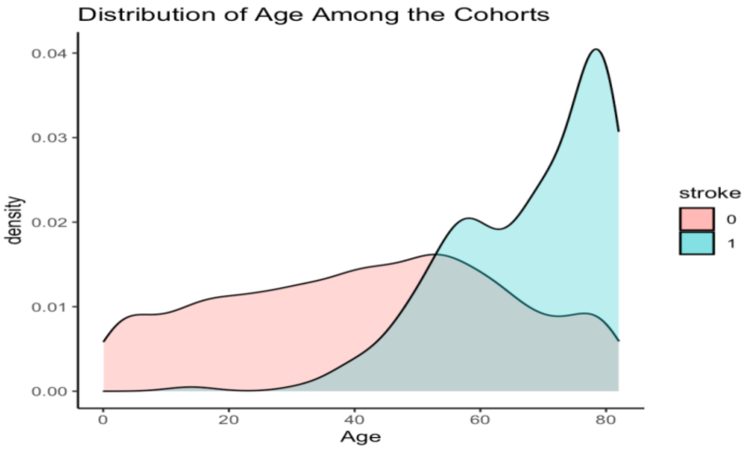


Figure 2

Age

- Figure 2 shows us the distribution of Age across different cohorts
- The distribution of non-stroke members is almost evenly distributed and does not change much.
- The stroke member distribution follows a left-skew distribution and provides a very good indication that the occurrence of stroke dramatically increases around 50. It also indicates that the risk of getting stroke increases with age.

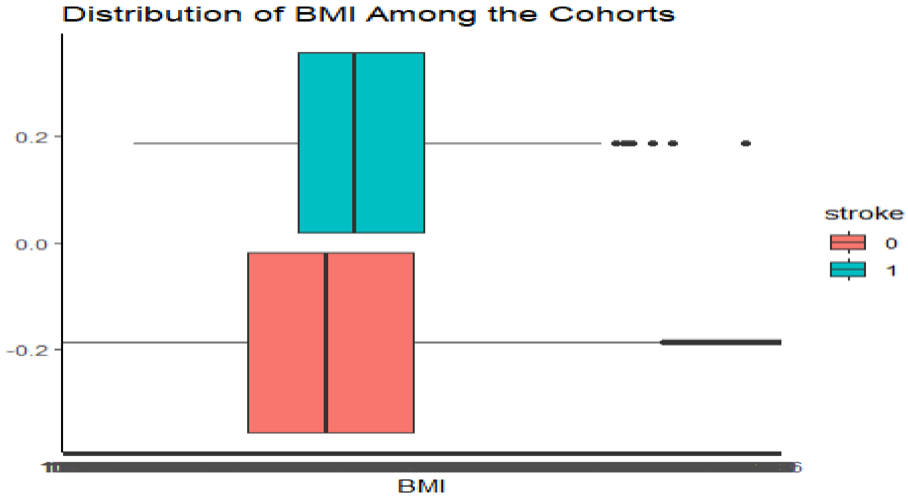


Figure 3

BMI

- Figure1 shows us the distribution of BMI across different cohorts
- When comparing the stroke and non-stroke cohort, there is only a slight difference between the two distributions.
- The stroke cohort has a slightly higher BMI compared to the non-stroke cohort.
- This may not be a good factor to consider as there is no significant difference between the two cohorts.

Categorical Variables:

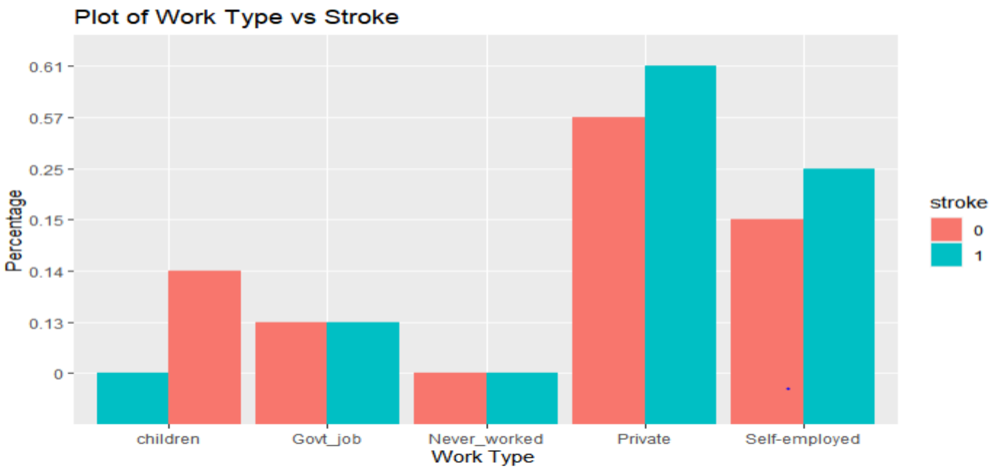


Figure 4

Figure 4 shows the type of work distributed by percentage for stroke and non-stroke members. The observations are as follows:

- There is a higher risk of getting stroke if working in Private or Self Employed.
- This insight could be biased as the people working are more likely to be around 25 or higher and could be correlated to age. Moreover, there is not enough data on never worked segment.
- Also, as children never work and are young, it skews the representation.
- This may or may not be a good variable to detect stroke.

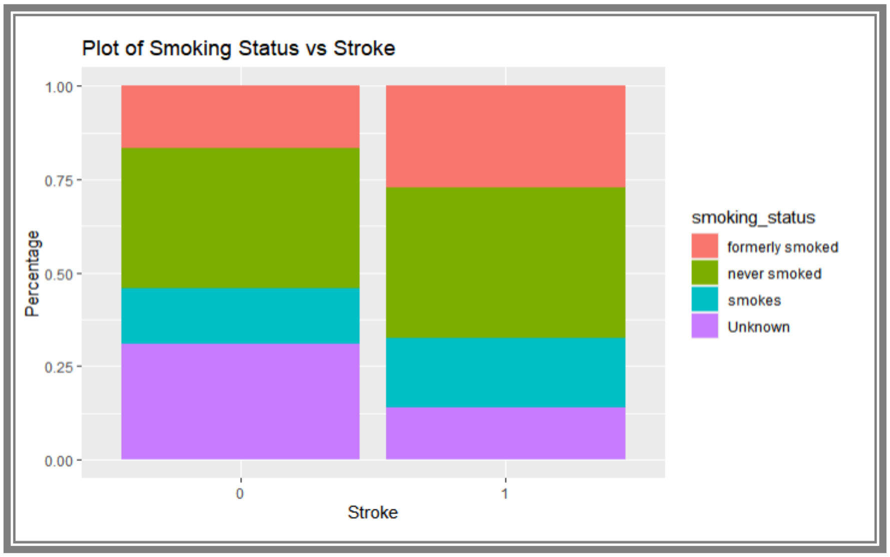


Figure 5

Figure 5 shows the smoking status distributed by percentage for stroke and non-stroke members. The observations are as follows:

- There is a higher risk of getting stroke if the person has formerly smoked.
- The rest of the segments show similar percentage distribution except the Unknown segment. Since the Unknown does not contribute much, this can be ignored.
- This variable could be a good indicator for the detection of stroke.

Bivariate Analysis

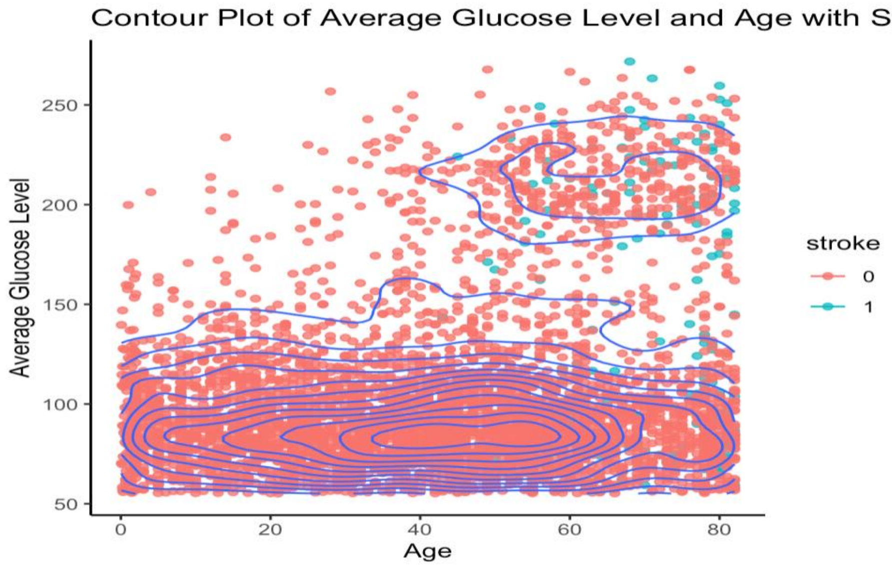


Figure 6

Figure 6 shows the scatter and contour plot of Average Glucose Level and Age for stroke and non-stroke members.

The observations are as follows:

- There is formation of two distinct clusters with the top cluster having more stroke members indicating that there could be some hidden properties which is being captured.
- There are hardly any stroke members in the lower cluster and the ones which are present are in the peripheral parts of the lower contour plot.
- We can interpret that most people who didn't have stroke had a lower average glucose level across all ages.
- This affirms our hypothesis that higher average glucose level and the older a person is, the higher is the risk of stroke.

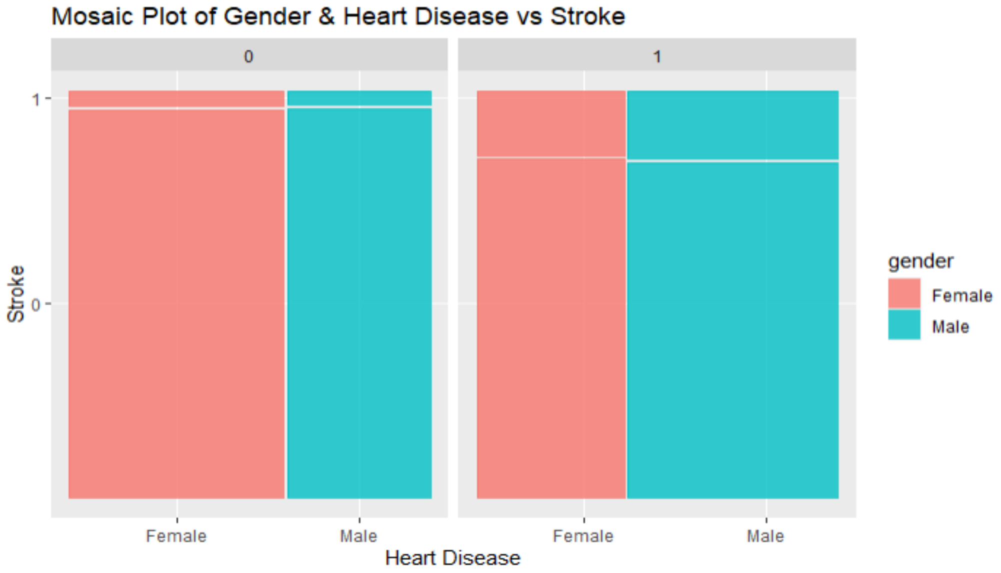


Figure 7

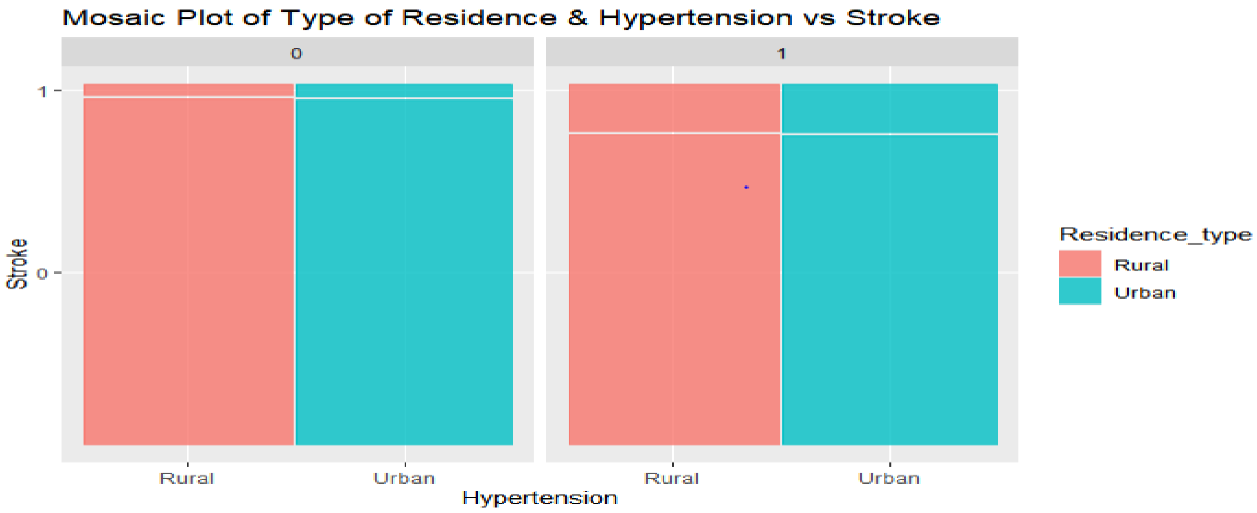


Figure 8

Figures 7 and 8 show the mosaic plots of two variables for stroke and non-stroke members. The left graph shows the mosaic plot of Gender and Heart disease, and the right graph shows the mosaic plot of Type of Residence and Hypertension for the two cohorts. The following observations can be made:

- The gender and type of residence hardly show any significant variance suggesting that they do not contribute much to the detection of stroke.
- For the heart disease variable, there is a significant difference and there is a higher chance of getting a stroke if there is a history of heart disease. A similar observation can be seen for the Hypertension variable too although there is lesser significance compared to heart disease.
- These two variables could slightly contribute towards the detection of stroke.

On investigation of the ever_married variable, it could be seen that there was some significance but as it could be correlated with age, further exploration was done by creating age brackets and building a mosaic plot of the age brackets and ever_married for the stroke and non-stroke cohort as shown below:

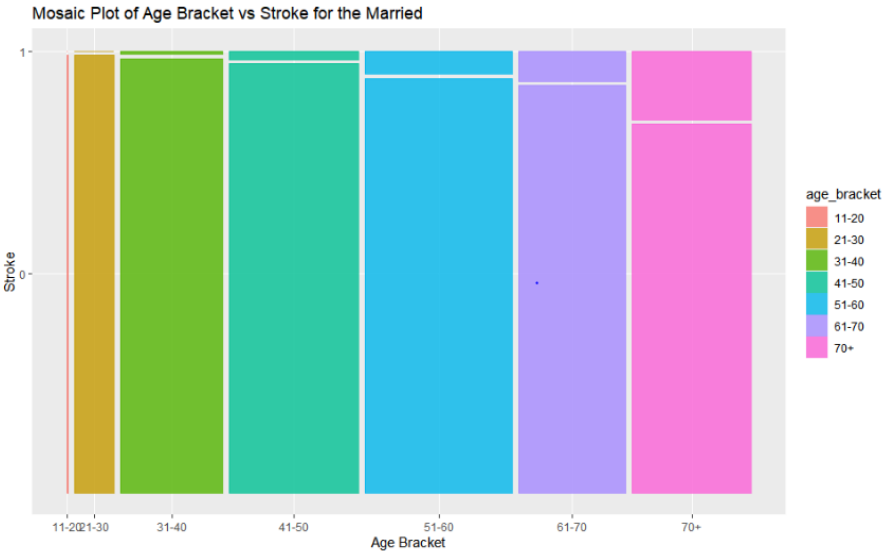


Figure 9

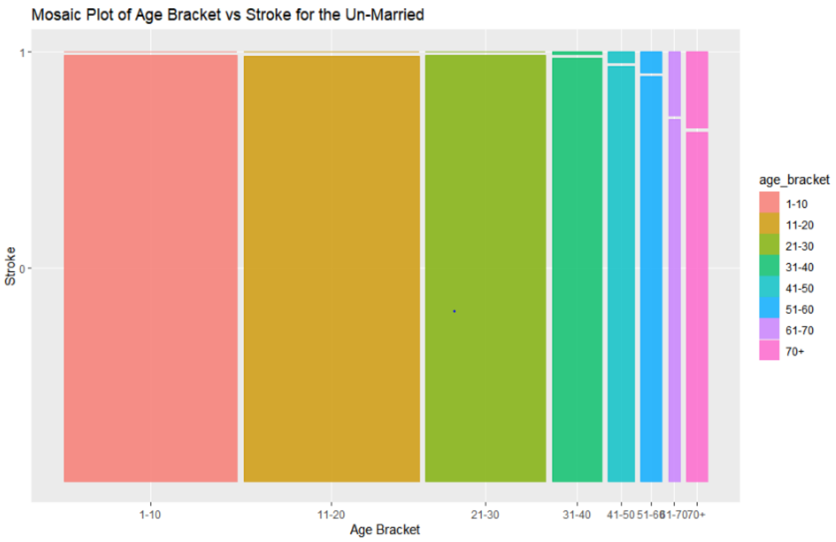


Figure 10

Figures 9 and 10 show if there is a relationship between ever_married and stroke
The observations are as follows:

- Thedistribution of members in the brackets for married and un-married members. This is observed as those who are married are only above 20 years and there is lower probability of members not being married after a certain age.
- A key thing to be noticed here is that the there is an equivalent rise in the risk of stroke for both the married and un-married cohort members which shows that the age variable is the one influencing it.
- The age variable is the key variable to consider in the detection of a stroke but the ever_married variable does not show much significance.

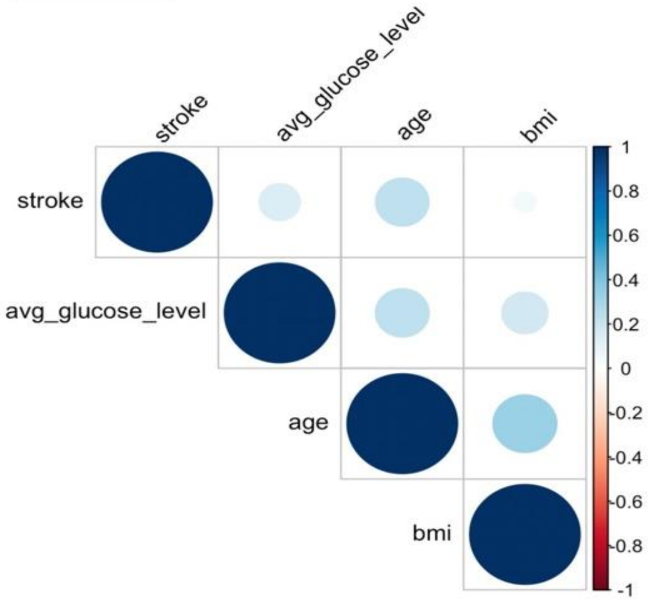
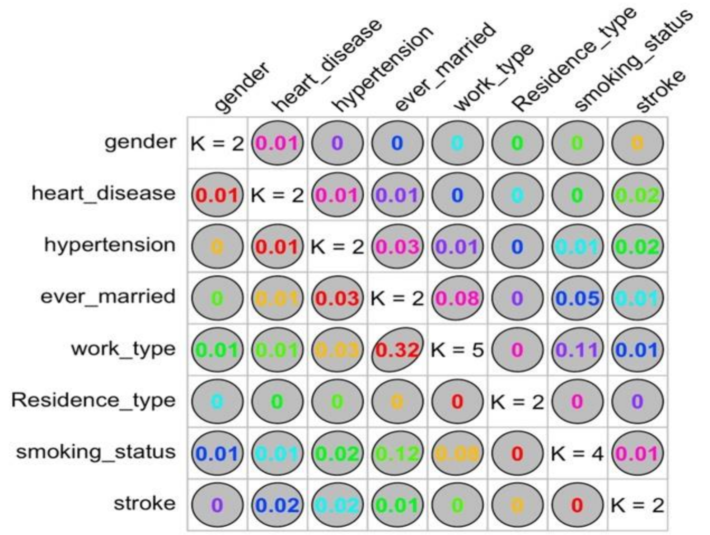


Figure 11

- In the graph described in Figure 11 we use the Pearson correlation to calculate the association between continuous variables.

- From this graph we can observe that Age and stroke are positively correlated as proved by our previous analysis in Figure 2
- Average glucose level is also slightly positively correlated with strokeproved in the analysis in Figure 1
- We also observe that Average glucose is slightly positively correlated with age proved in the analysis of Figure 6
- BMI seems to be the least positively correlated with stroke depicted in the analysis of Figure 3



	gender	heart_disease	hypertension	ever_married	work_type	Residence_type	smoking_status	stroke
gender	K = 2	0.01	0	0	0	0	0	0
heart_disease	0.01	K = 2	0.01	0.01	0	0	0	0.02
hypertension	0	0.01	K = 2	0.03	0.01	0	0.01	0.02
ever_married	0	0.01	0.03	K = 2	0.08	0	0.05	0.01
work_type	0.01	0.01	0.03	0.32	K = 5	0	0.11	0.01
Residence_type	0	0	0	0	0	K = 2	0	0
smoking_status	0.01	0.01	0.02	0.12	0.08	0	K = 4	0.01
stroke	0	0.02	0.02	0.01	0	0	0	K = 2

Figure 12

- Since our dataset has many categorical variables in Figure 12, we use the Goodman Kruskal tau measure to calculate the association between categorical variables.
- The Goodman and Kruskal tau measure is an asymmetric association measure between two categorical variables, based on the extent to which variation in one variable can be explained by the other.
- The diagonals depict the number of levels or categories available in each variable.
- From Figure 12 we interpret that no variable is perfectly predictable from any another variable.
- The reasonable association seen in this plot is that between work_type and ever_married where the forward association is 0.32 and the reverse association is 0.08 which is almost negligible predictive power.

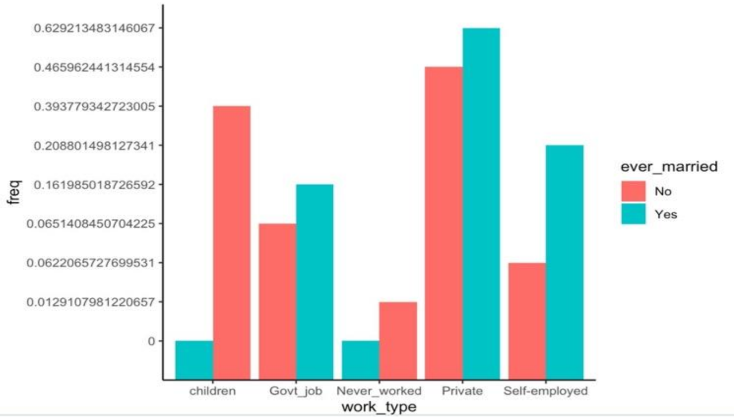


Figure 13

- From the Figure 13 we show proof of concept that there is one way association between work_type and ever married
- We can observe that children and never_worked were never married. Which means that given the work_type children or never_worked we would be able to predict that they were never married. But given that they were never_married we cannot predict their work_type

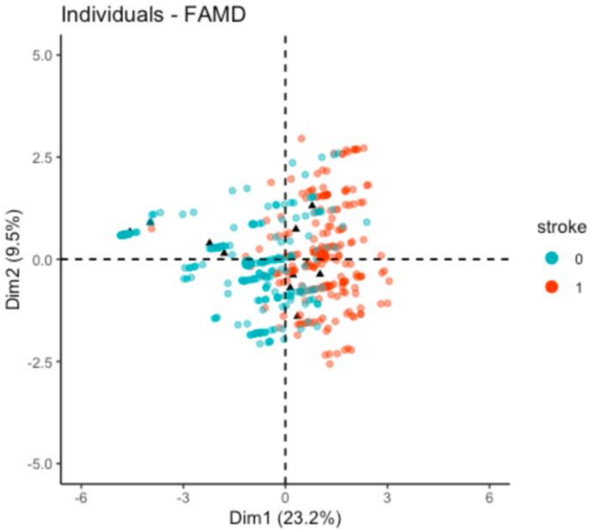
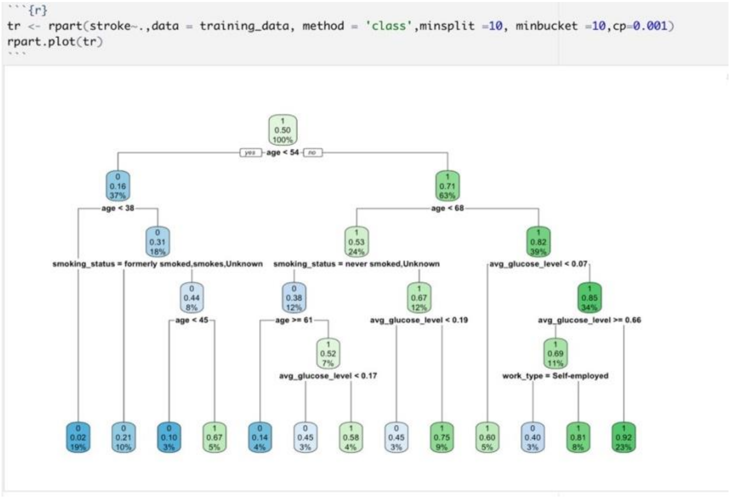


Figure 14

- Figure 14 explains the Factor analysis of mixed data (FAMD) is a principal component method dedicated to analyzing the data set containing both quantitative and qualitative variables. It makes it possible to analyze the similarity between individuals by considering a mixed type of variables. Additionally, one can explore the association between all variables, both quantitative and qualitative variables.
- The FAMD analysis in Figure 14 is done after balancing the data. We take 209 datapoints from the stroke and the non-stroke cohort respectively.
- From this figure we can understand that only a small percent of variation in the graph is represented by the first 2 principal components.
- This suggests that there are not a lot of correlation among the variables, so we cannot significantly reduce the dimensions.
- Hence, we move on to non-linear classifiers.



~~~
```{r}
tr <- rpart(stroke~.,data = training_data, method = 'class',minsplit =10, minbucket =10,cp=0.001)
rpart.plot(tr)
```
~~~

Figure 15

- In Figure 15 we use a decision tree as our non-linear classifier. We have 50 percent of patients not having strokei.e. we run our experiments on a balanced dataset.
- We used the following hyperparameters for the tree:
    - Minsplit=10
    - Minbucket=10
    - Cp=0.001
- We can interpret the tree as follows:
    - The root node asks whether the age is less than 54. If it does, then the probability of having stroke is 16 percent.
    - If it doesn't, then we go down to the root's right child node. 63 percent are patients with age larger or equal to 54, and the probability of having stroke is 0.71
    - We keep on going down like that to understand how features impact the likelihood of having stroke.

Conclusion:

From the above analysis, we can conclude the following:

- The variables Age, Average Glucose level, Smoking Status and Work Type showed a good significance in the detection of stroke.
- Although, the variable ever_married initially showed significance in our analysis, the significance was attributed by age and not ever_married.
- The variables are highly uncorrelated and thus, the features cannot be reduced using PCA or other dimensional reduction techniques.
- The non-linear classifier 'tree', provided good significant results as it was able to extract the important features which were previously hypothesized to be significant.

Limitations:

- One of the primary limitations we faced was dealing with a highly imbalanced data.
- There could be other key factors affecting the stroke like cholesterol.
- The data might not be a good representative of the population.

Future Scope:

- Better data acquisition such as a balanced data with other key features can help in giving better insights such as Blood Pressure, Cholesterol and other more important factors related to health conditions.
- Implementation of Stroke analyses in real-time environments.
- Determining the severity of the stroke.