Stroke prediction model

Team Members:

1- Yehia Hamdy 2018170936
2- Mazen Amer 2018170930

Project Description:

In this project we analyze a dataset, and the main factors that cause the attribute class to happen. It also visualizes the data by showing the data spread by using charts and graphs. After we apply Decision tree algorithm, SVM, and many others to extract rules from our dataset. Finally, we can predict the probability of A patient having a stroke.

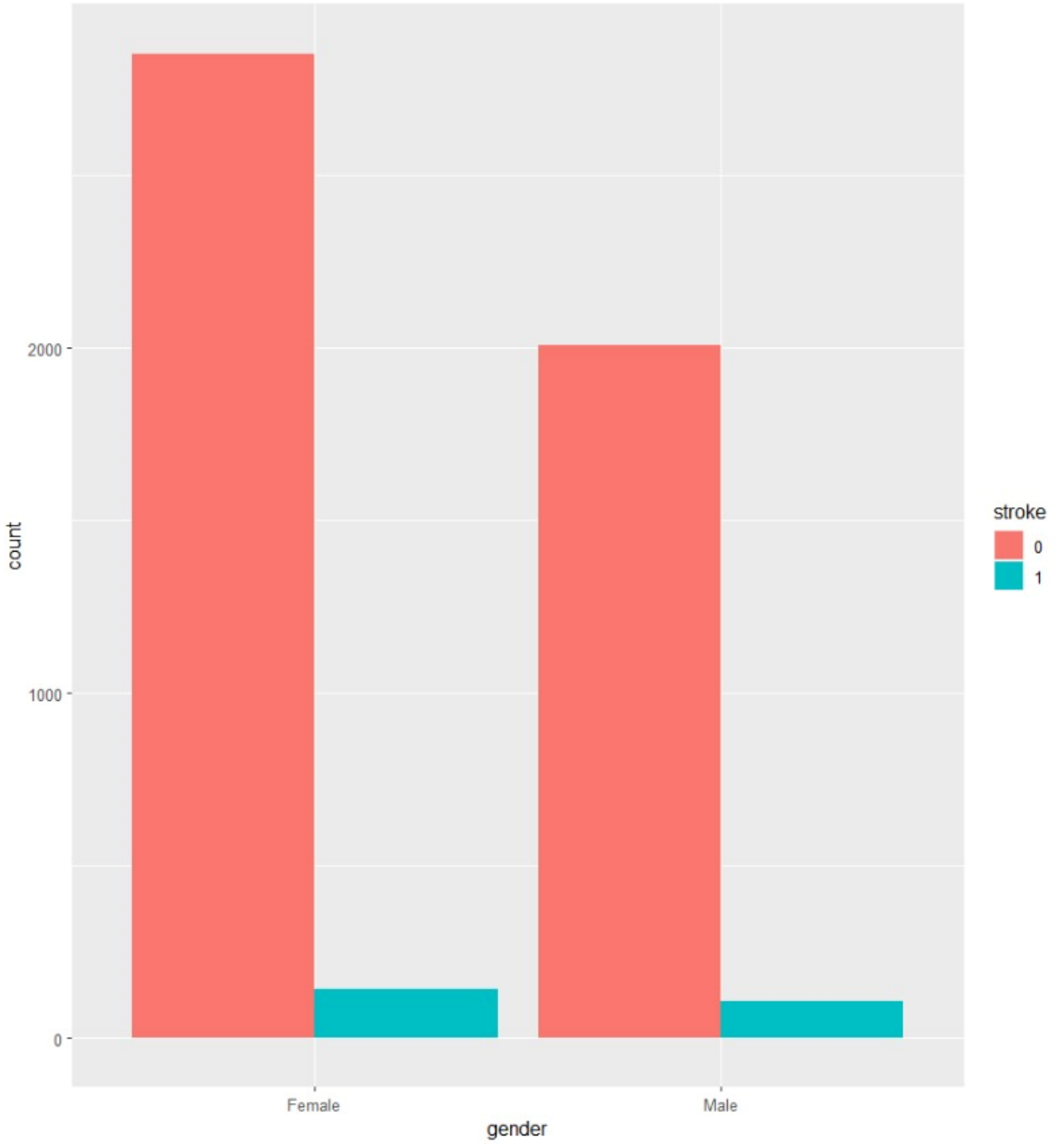
Data-set Brief Description:

According to World Health Organization, stroke is the 2nd leading cause of death globally which approximates to 11% of total deaths. Study shows that strokes can be prevented 80% of the time, through proper education, on signs of the stroke. Hence, it's critical that hospitals actively participate in Stroke prediction to eliminate treatment delays and improve outcomes for new and acute stroke patients.

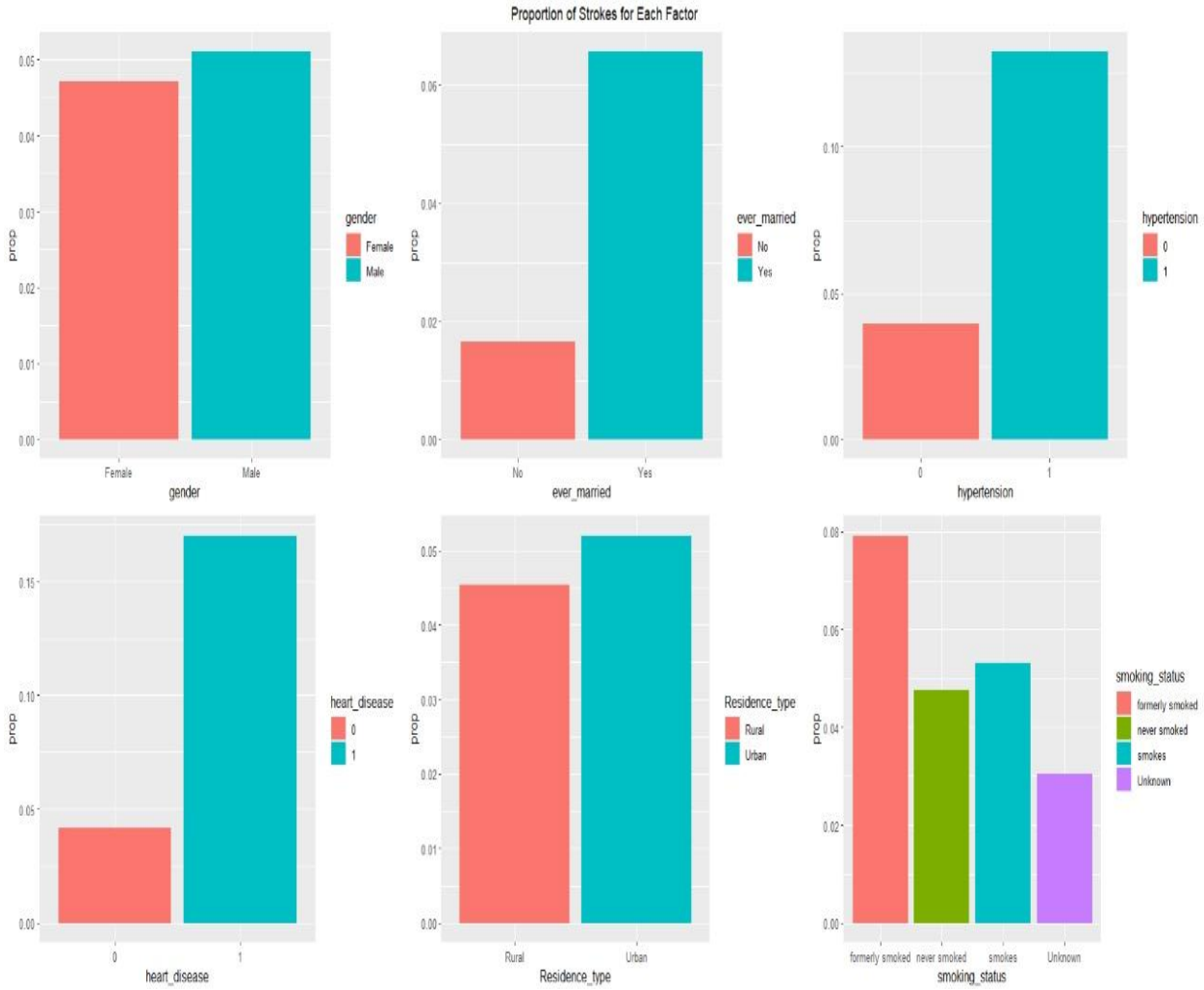
problem definition and objectives:

The problem in our case is that we have raw data containing a lot of knowledge with no way to extract that knowledge. Our objective is to find a way to collect that knowledge by using big data technics so that we could use this data in helping doctors predict weather or not his patient will have a stroke.

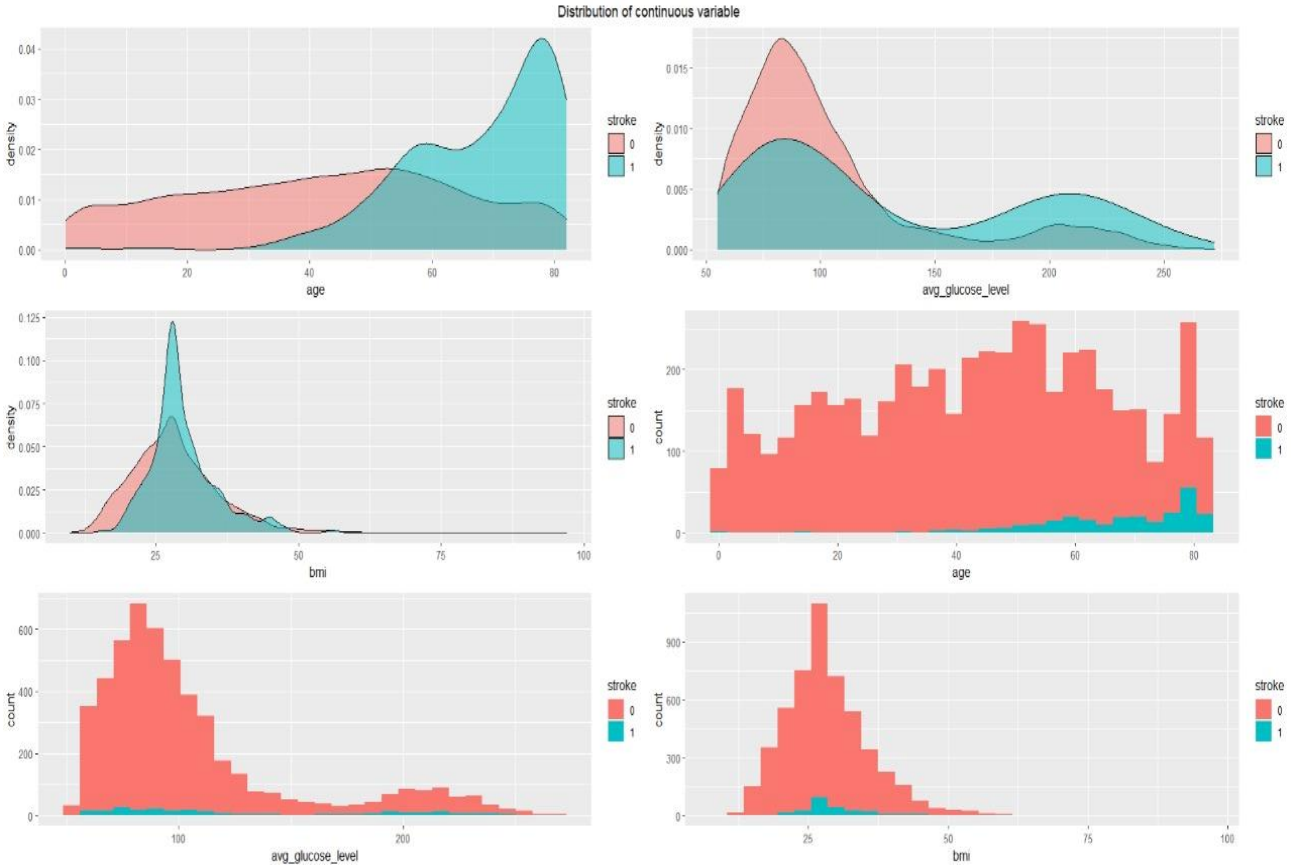
Screenshots of the data visualization charts and your observations on each chart:



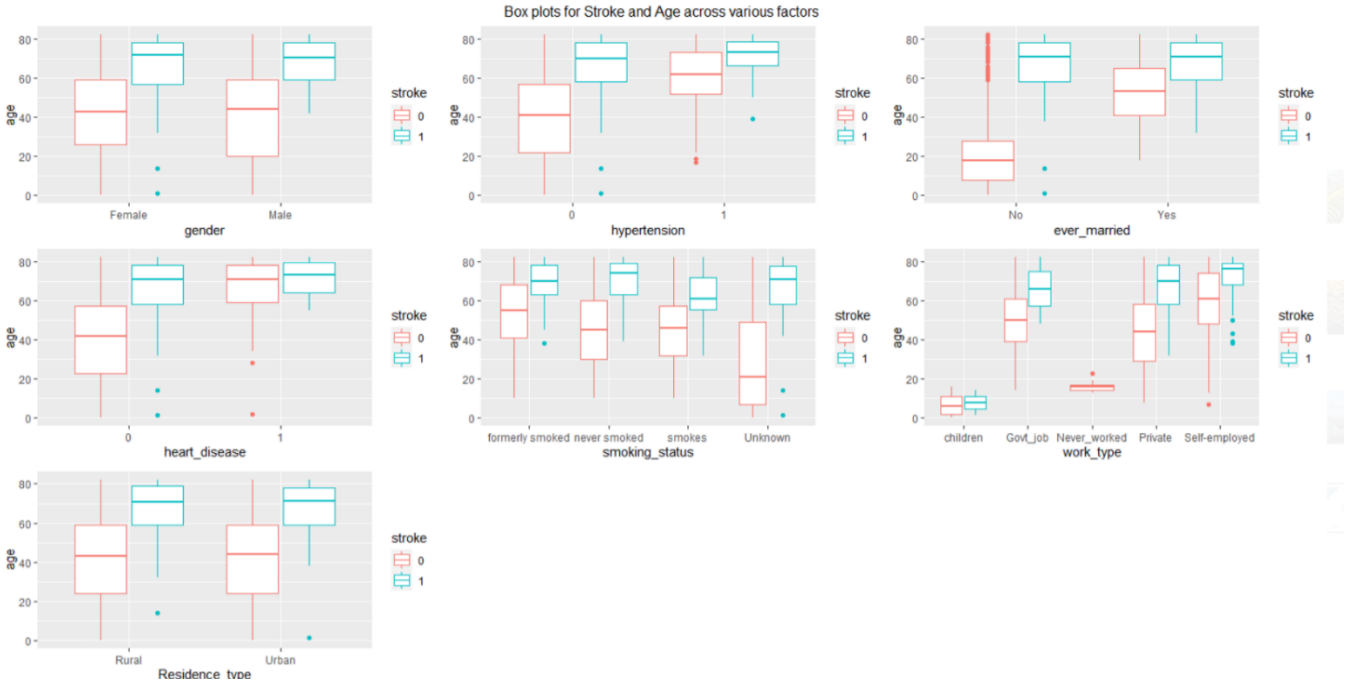
This graph shows that the dataset has more females than males in general



Because we just saw that the dataset isn't even. So, we calculated the probability to even out the data and the graph above shows the probability of every factor related to the person having a stroke or not. First the graph shows that males have a higher chance of getting a stroke than females. Second married people have a higher chance of getting stroke than non-married people. People with previous heart disease or hypertension get stroke more. You haver a higher chance of getting a stroke if you live in urban areas, but it doesn't differ that much from rural areas. People who are smokes also has a greater percentage of getting the stroke when compared to non-smokers.



In this graphs we show the count and the density of three factors (age, BMI, glucose level). We can see that the number of stroke increases as the age increases. The glucose levels is bimodal, so when the glucose is higher or lower than 150 there is a bigger chance of getting a stroke. And the BMI doesn't affect the chance of getting a stroke.



Most of the people who got stroke are the people who are older. People who are self employed are also older than rest of the population. People who had stroke and smokes are younger than those who never smoked.

The applied data cleaning or transformation methods:

We removed from the gender field “Others” as there was only one patient. Also, we removed all rows that contained Null values as there was only around 200 so it didn’t affect the size of our dataset.

The dataset preparation in terms of machine learning (training set, learning set):

We set the training to be 80% of the data, The rest of data we set to be testing (20%).

The used data analytics techniques:

- Random Forest
- when we have a large dataset, and interpretability is not a major concern.
- Logistic Regression
    - logistic regression is also used to estimate the relationship between a dependent variable and one or more independent variables
- Support Vector Machine
    - SVM is used when a number of features are high compared to a number of data points in the dataset
- Decision Tree
    - Decision trees are used for handling non-linear data sets effectively
- Best Pruned Tree
    - Pruning is a data compression technique that reduces the size of decision trees by removing sections of the tree that are non-critical

The performance measures used to evaluate your data analytics technique and your analysis/discussion for the whole project findings:

Model	Accuracy	Sensitivity
Logistic Regression	87.21	88.90
Random Forest	92.71	96.49
Best Pruned tree	88.35	91.10
Decision Tree	83.39	84.84
SVM	81.42	83.08