Final Report:
Stroke Prediction Model

Problem Statement:

Stroke is the sudden death of some brain cells due to lack of oxygen when the blood flow to the brain is lost by blockage or rupture of an artery to the brain. According to the World Health Organization (WHO) stroke is the 3rd leading cause of death and 2nd leading cause of disability worldwide. It is also a leading cause of dementia and depression. According to the WHO stroke is responsible for 11% of total deaths world wide.

Identifying if a person can get a stroke will help the health care professionals to come up and suggest effective prevention strategies including lifestyle changes. WHO studies have found that preventive strategies have proved effective in reducing stroke mortality.

Dataset:

The data set is collected from Kaggle as a CSV file.
The data contains physiological and biological information about patients like, Average Glucose Level, Gender, Age, BMI. It contains patients diseases history like, hypertension, heart disease, stroke. It also contains the patient's other information such as work type, residence type, smoking habit, and marital status.

Overview:

The prediction model was build using following steps:

1. Data Wrangling
2. Exploratory Data Analysis
3. Data Preprocessing
4. Modeling

Programming language and tools:
Python
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn
- Scipy

1. Data Wrangling

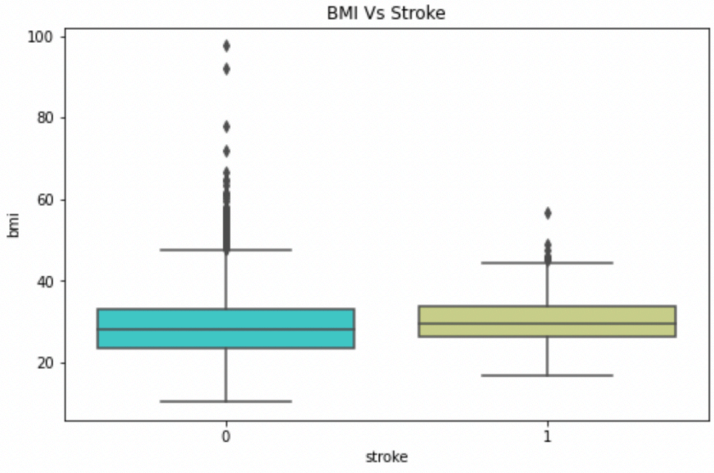
I downloaded the data set from Kaggle in CSV format.
I loaded the data using python pandas in a data frame and inspected the data.
The raw dataset had 5110 entries and 12 features(columns).
The dataset had both discrete and categorical features.

```
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
```

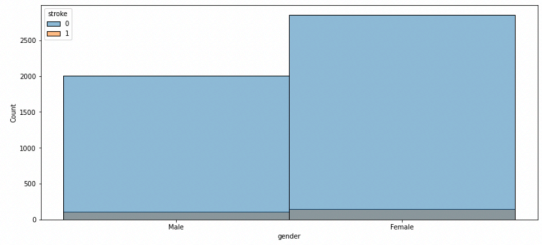
I checked the data for null or missing values and observed that the ‘bmi’ column had 201 missing values. I did the point inspection on rows corresponding to the ‘bmi’ column with null values. I also found the BMI column to have some outliers.

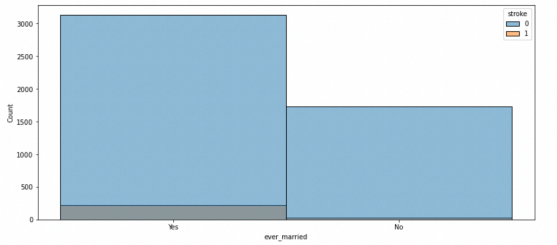
I inspected the data in the ‘gender’ column. It had 3 unique values ‘Male’, ‘Female’, ‘Other’. In taking the count on the number of records for each unique value I found that there was only 1 record for ‘Other’. I deleted this record since it would not have been of any significance on the outcome of the model.

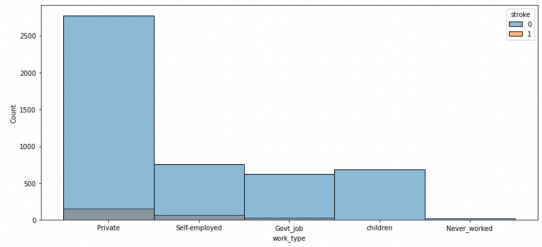
I checked the data for the dependency. I checked the dependent feature ‘stroke’ for its dependency/relation with the independent feature ‘bmi’, and I could not find any strong correlation was between the two. I also checked for the dependency between the dependent feature ‘stroke’ and the other independent categorical features in the dataset using different plots such as box plot and histograms. Features like, gender, work_type, residence_type were found to show some correlation.

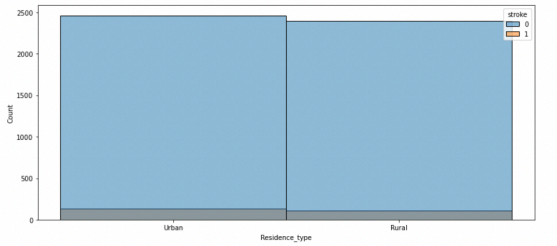


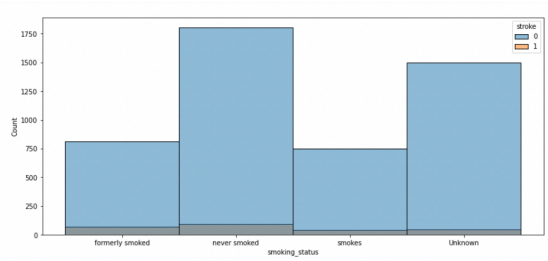
Plots of independent categorical features and dependent feature 'stroke' to check the correlation.











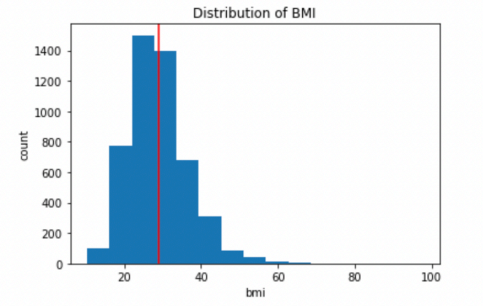
I saved the csv file in 'data' folder to be used in the next steps of Exploratory Data Analysis and Processing.

2. Exploratory Data Analysis

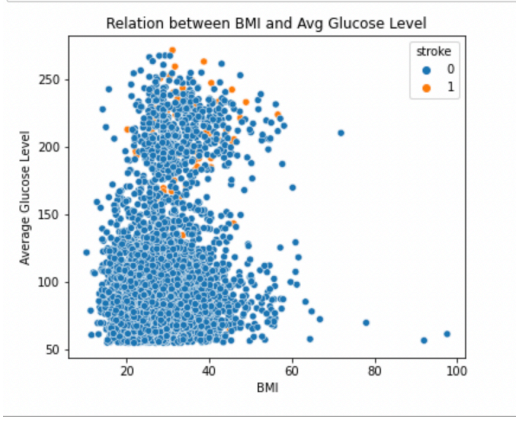
The data set had 12 columns and 5109 rows from the Data Wrangling step. The features in the data are mostly categorical except for 'age', 'bmi' and ‘average_glucose_level’ which are the only numeric features.

I dropped the id column and performed the visual exploration on 11 columns and 5019 rows.

I checked if any columns have missing values and found that the 'bmi' column which had some missing values. I checked for the distribution of the data in the ‘bmi' column and observed that the values were distributed around the mean. I performed the imputation on the ‘bmi’ column. I replaced the missing values with the mean of the data in the column.

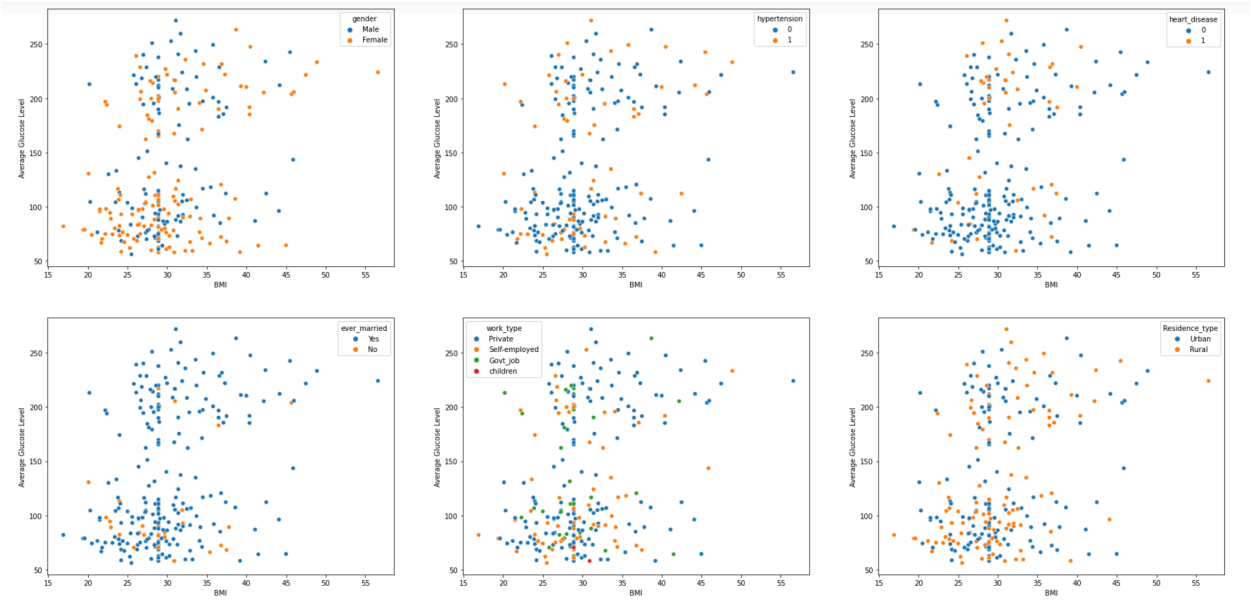


I analyzed the correlation between the combination of features, such as, Average Glucose Level, BMI and the dependent feature stroke using visual analysis. I observed a strong correlations between these features. It is observed that the people between the age group of 60-80 are at higher risk of getting stoke. The risk is even higher for the people over 80years of age. I also observed that the people with BMI between 30-60 along with the Average Glucose Level of more than 130 are also at higher risk of getting stroke.

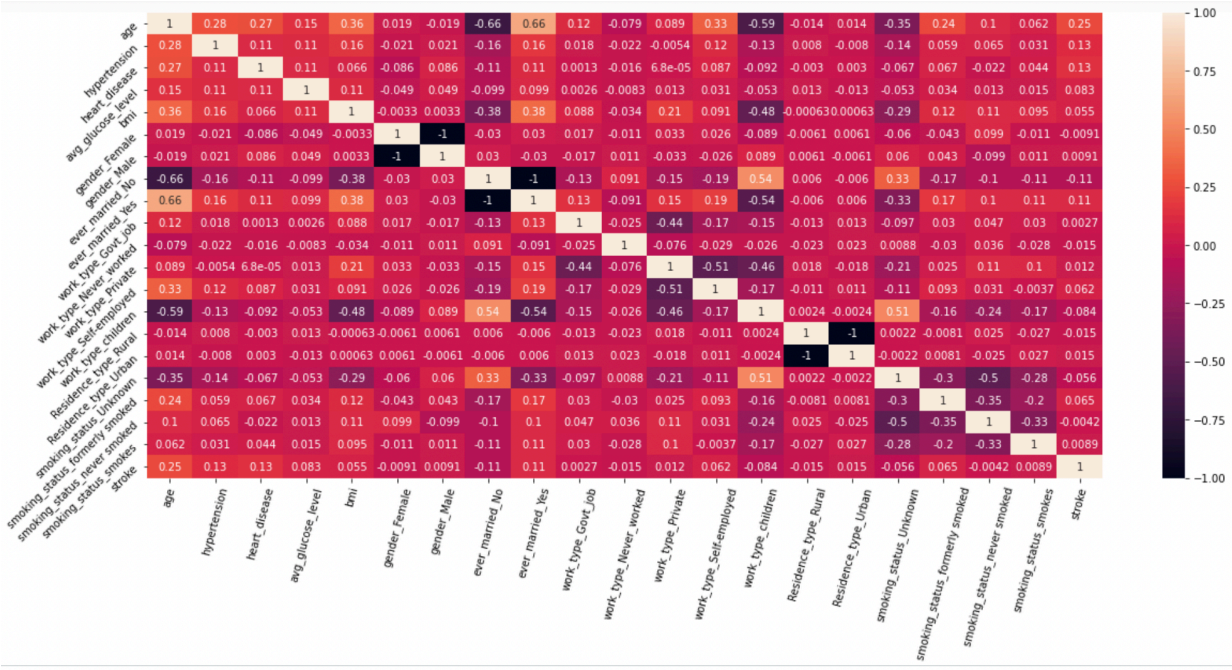


I made following observation while performing the visual analysis for correlation between the independent features:

- Average Glucose level is higher in men as compared to women with similar BMI.
- There are more people suffering from Hypertension with BMI ranging between 25-40 with Average Glucose level higher than 150
- There are more people suffering from Heart Diseases with BMI ranging between 25-40 with Average Glucose level higher than 200
- People working in Private jobs are having higher Average Glucose Level compared to people who are self employed, having government jobs.
- People living in rural areas are having lower Average Glucose Level and BMI lower than 30
- People with smoking history or current smokers have relatively higher BMIs.



I checked for the strength and the direction of the relationship between the features using Spearman Rank Correlation Coefficient but could not observe strong correlation between individual independent variable and the dependent variable.



Finishing the Exploratory Data Analysis step I concluded that
The dependent feature stroke seems to have strong correlation or dependency on the combination of independent features like

- Higher BMI along with - hypertension - heart disease - marital status - smoking habits
- Average glucose level with - hypertension - heart disease - marital status - smoking habits

Imputed data set was saved in the ‘data’ folder for the next step.

3. Data Preprocessing

I used the data set from the stored CSV file at the end of the EDA. The data had 12 columns and 5109 rows.

I checked the unique values for the categorical data. The unique values for the categorical data were as follows

gender -> 'Male', 'Female'
ever_married -> 'Yes', 'No'
work_type -> 'Private', 'Self-employed','Govt_job', 'children', 'Never_worked'
Residence_type -> 'Urban', 'Rural'
smoking_status -> 'formerly smoked', 'never smoked', 'smokes', 'Unknown'

I converted all of the categorical independent variables to dummy discrete variables(0 or 1 values). This increased the width of the dataset from 12 columns to 21 columns.

I further analyzed the data for the class imbalance and observed that out of 5019 records 4820 rows were for ‘stroke = 0’(no stroke) and only 249 records for ‘stroke = 1’. Since this imbalance of data would have lead to biased model behavior, I up-sampled the data. I created more rows of data for the value of ‘stroke = 1’.

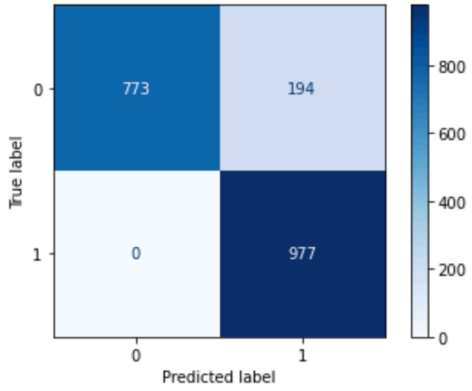
The data was then split into 80/20 ratio for training and testing.
I performed data scaling to bring the data of all of the features on to the same scale.

4. Modeling

I modeled the data using classification machine learning algorithm to predict the risk of stroke depending on the medical and physiological conditions.
I used 4 different types of classification models and I compare the the performance of all the models to identify the best model to predict the risk of the stroke.

Model 1: KNeighborsClassifier using Grid Search CV for Hyperparameter optimization

```
Testing Score :  0.9002057613168725
Confusion Matrix
```




```
0 - No Risk Of Stroke(Positive)
1 - High Risk Of Stroke(Negative)
```

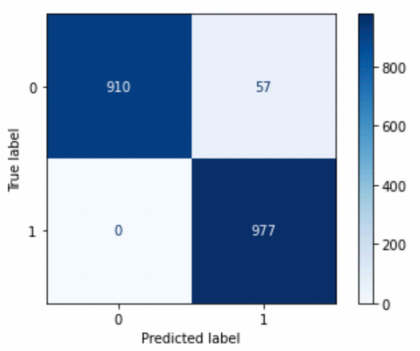
I used Grid Search CV for parameter optimization.The best parameter obtained by implementing GridSearchCV was n_neighbors = 8.

I used n_neighbors = 8 on KNeighborsClassifier model, trained the model using the training set and tested the model using the test set. The model gave a test score of 90%

Out of 967 actual positive(no risk) values:
The model made 773 correct positive(No Risk of Stroke) predictions.
194 were inaccurately predicted as negative(High Risk Of stroke)
Out of 977 actual negative(high risk) values:
The model made 977 correct negative(high Risk of Stroke) predictions with no inaccurate predictions for (No risk of stroke)

Model 2: Decision Tree

```
The test score for DecisionTree is : 0.9706
```



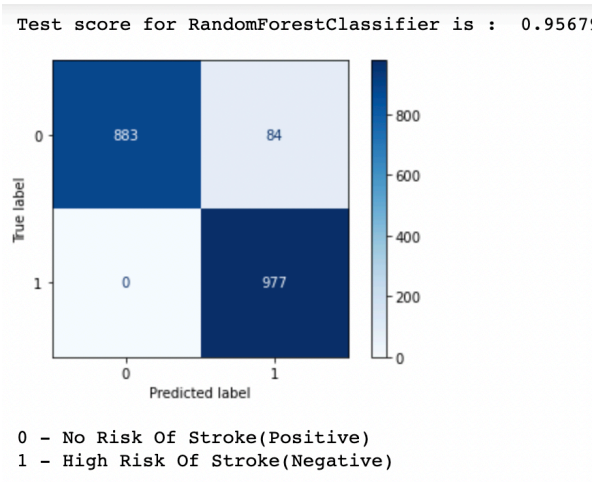

```
0 - No Risk Of Stroke(Positive)
1 - High Risk Of Stroke(Negative)
```

I trained and tested the Decision Tress classification algorithm using the train and test set
The model gave a test score of 97%

Out of 967 actual positive(no risk) values:
The model made 910 correct positive(No Risk of Stroke) predictions.
57 were inaccurately predicted as negative(High Risk Of stroke)

Out of 977 actual negative(high risk) values:
The model made 977 correct negative(high Risk of Stroke) predictions with no inaccurate predictions for (No risk of stroke)

Model 3: RandomForestClassifier using RandomSearchCV for Hyperparameter optimization



I used RandomSearchCV for hyperparameter optimization.
The best parameters I obtained were max_depth=46, min_samples_leaf=4, n_estimators=1800.

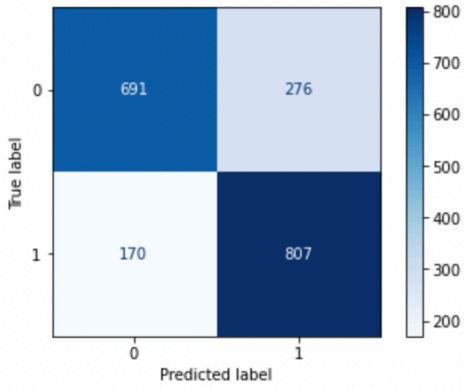
Using these hyperparameters I trained and tested the model using train and test data. The model gave a test score of 95%

Out of 967 actual positive(no risk) values:
The model made 883 correct positive(No Risk of Stroke) predictions.
84 were inaccurately predicted as negative(High Risk Of stroke)

Out of 977 actual negative(high risk) values:
The model made 977 correct negative(high Risk of Stroke) predictions with no inaccurate predictions for (No risk of stroke)

Model 4: LogisticRegression

```
Testing Score for LogisticRegression is :  0.7705
```



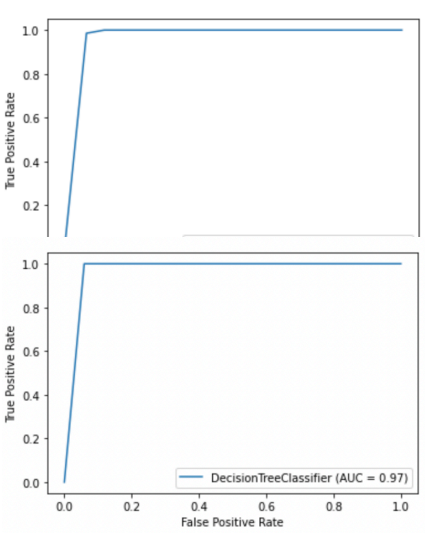

```
0 - No Risk Of Stroke(Positive)
1 - High Risk Of Stroke(Negative)
```

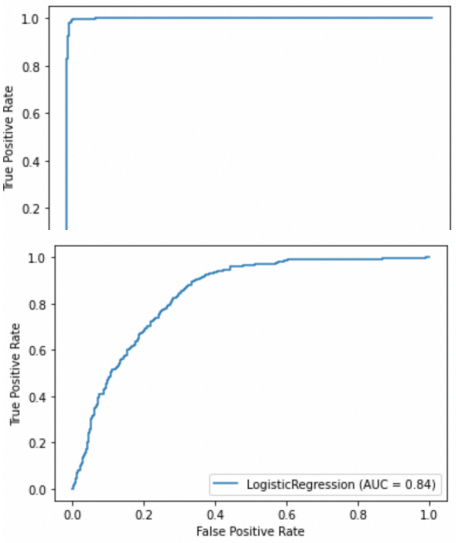
I trained and tested the Logistic Regression algorithm using the train and test data. The model gave a test score of 77%

Out of 967 actual positive(no risk) values:
The model made 691 correct positive(No Risk of Stroke) predictions.
276 were inaccurately predicted as negative(High Risk Of stroke)

Out of 977 actual negative(high risk) values:
The model made 807 correct negative(high Risk of Stroke) predictions.
170 were inaccurately predicted as negative(Low Risk Of stroke)

5. Conclusion





I compare the performance of all the models using the ROC curve and the classification report and confusion Matrics. I did not select the model on the basis of best test score but on the basis of how generalized the model is. I also considered the sensitivity of the model. Since this model is for predicting the people who are at higher risk of getting stroke, it is important that model should not erroneously predict the patient at actual risk as not a high risk patient. It is safer to predict patient as high risk erroneously even if they are not at high risk.

I concluded that RandomForest with an overall prediction accuracy of 95% is the best model. The model is balanced and is likely to make accurate predictions for positive values (people with a low risk of getting a stroke) and not identifying 'people with high risk' falsely to not be at risk of getting a stroke. The model is sensitive to negative values, and for the problem at hand; having a false negative is a superior alternative to having a false positive

6. Future Scope

This model is helpful for health care practitioner for predicting if the patient is at high risk of getting a stroke depending on the physiological parameters and medical history. This will help the practitioner plan the preventive treatment/strategy for their patients.

I would like to expand this model further to predict the high risk of death by stroke by adding more features.

The data set doesn’t have geographical and economical date. I feel that both of these factors might be of significance because the economic status, surrounding environment, and the availability of resources largely effects a persons lifestyle thus effecting the outcome of a disease.

For example a person with the medical condition such as heart disease, hypertension or diabetes living in a developed nation and with higher economic status is more likely to have resources such as medications and other help to avoid the risk of getting a stroke and dying. On the contrary person having similar medical conditions living in underdeveloped or poor economies will less likely to get help or resources and thus will be at higher risk of dying if he gets stroke.

I would like to consider factors like country of residence, economic status, number of deaths by stroke in the country along with physiological and biological factors. This will make the model generalized for the worldwide population and can be used by the authorities to to focus on planning the preventive strategies especially for the poorer nations.