Project Proposal – Progress Report

# Initial Approach

The real-world problem that will be solved with this problem is trying to predict whether a patient will have a stroke in their future, based on qualities like BMI, age, gender, smoking habits etc. For this problem, the data set was collected from Kaggle. The data contains 12 columns and 5110 rows of data, there will be 11 variables in total and the predicted variable will be a variable ‘stroke’ which is a binary classification variable that indicates whether the patient has had stroke or not, indicated by 1 and 0 respectively.

For evaluating the model, recall will be used as the primary reference when comparing models. Accurate classification and reasonable intervention for high-risk population can effectively reduce the burden of stroke on families and the society. It is necessary to consider the recall of the classification model to ensure the pertinence of stroke intervention.

Graphical user interface, application, table, Excel

Description automatically generated

**Figure 1**

Above is a snippet of the data set, as we can see there will be 11 variables in total that will be predicting the variable “stroke”. Stroke is a binary variable indicated in 0 and 1. The 11 variables are as follows:

1. ID – id of the patient
2. Gender – gender of the patient
3. Age – age of the patient
4. Hypertension – if the patient has hypertension
5. Heart Disease – heart disease patient or not
6. Ever Married – married or not
7. Work Type – work type
8. Residence Type – residence type
9. Avg\_glucose – glucose level of patient
10. BMI – Body Mass Index of patient
11. Smoking Status – whether patient smokes or not.

# Current Approach and Work in Progress:

## Data Preprocessing

To start the data preprocessing, the ID column was dropped first from the analysis as it didn’t add any extra value to the current project. After that all the columns were checked for null values. The Figure below shows the count of the null values for each of the columns.

Graphical user interface, text, application

Description automatically generated

**Figure 2**

As we can see from the figure above, BMI had 201 null values in total, all the other columns have 0 null values. For the BMI column, the null values were imputed using mean. Once the null values were checked, variables were then categorized into numerical and categorical variables. The categorical and numerical variables are as follows:

|  |  |
| --- | --- |
| Numerical | Categorical |
| BMI | Gender |
| Age | Hypertension |
| Avg\_glucose\_level | Heart\_disease |
|  | Ever\_married |
|  | Work\_type |
|  | Residence\_Type |
|  | Smoking\_status |
|  | Stroke |

Using the describe function, the statistics of all the variables were checked to find any abnormalities in the data. From the gender variable, it showed 3 types of genders, Male, Female and Other. But for ‘Other’, there was only 1 instance of ‘Other’, I didn’t not want to classify this into the Male or Female category and since there was only 1 instance, I decided to leave it as it is.

Next, the EDA was completed for the variables to understand their distribution better.

A picture containing histogram

Description automatically generated

A picture containing logo

Description automatically generated

**Figure 3**

The figure above shows two images, the first from the top being the distribution of avg\_glucose\_level and the image below is the distribution of BMI. As we can see from the image, both variables are slightly skewed and there are numerous outliers. To deal with this, I did an outlier detection that removed any variables that was 1.5 \* IQR(Interquartile Range) of the lower and upper whisker. Once that was completed the variables did not seem as skewed and looked more normal, with little to no outliers, as seen from the Figure below.

Chart, box and whisker chart

Description automatically generated

**Figure 4**

EDA was conducted for the categorical variables as well, notably looking at the Figure below we can see that Stroke is heavily imbalanced, with only 4.9% of the total data showing patients that had stroke.

Chart, bar chart

Description automatically generated

**Figure 5**

My initial approach to deal with this situation was to use SMOTE to balance the dataset. However, after further research and understand, as SMOTE is outside the scope of this project, it wasn’t considered for the model building process.

A correlation plot was made between all the numerical variables to understand the relationship between the variables. As we can see from the Figure below, most of the variables don’t have a strong or notable relationship to one another. Only BMI and Age had a moderate relationship of 0.35.

Square

Description automatically generated with medium confidence

**Figure 6**

## Model Building

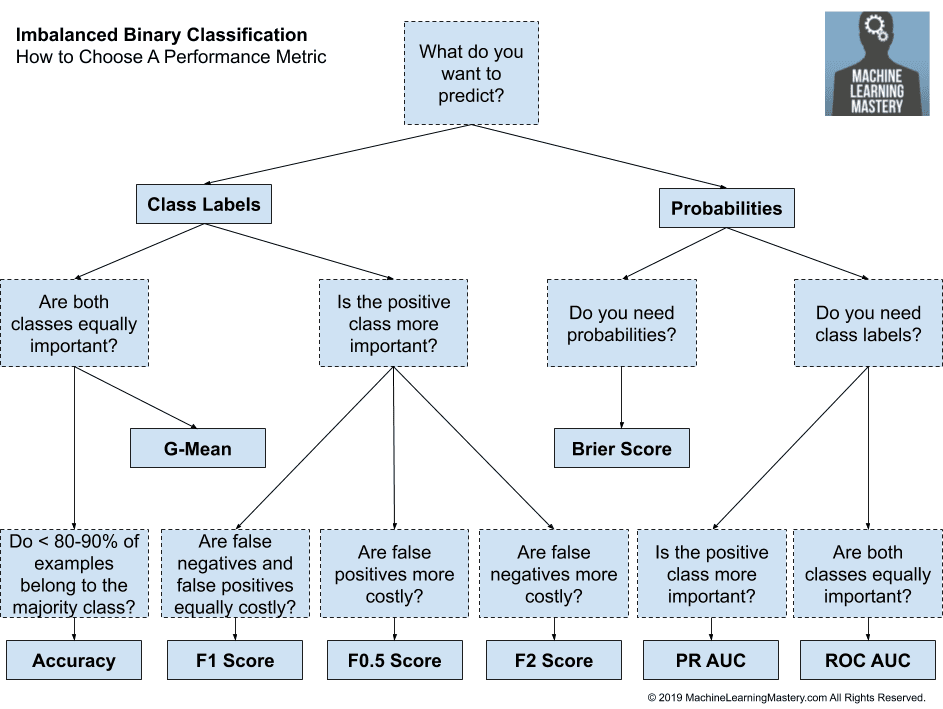
During my initial approach I decided to use Recall as the primary reference when comparing different models. The models that were initially planned for the project was:

1. Logistic regression
2. KNN
3. Decision trees
4. Random forests
5. SVM and Naïve Bayes

For the current approach, I first one hot encoded all the variables and then created dummy variables for each of the categorical variables, which increased the total number of variables to 16. I created models using some of the algorithms above to understand how they would behave with my current data set. The results were as follows for the testing data set:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | Accuracy training | Accuracy  Testing | Recall  Training | Recall  Testing | ROC  Training | ROC  Testing |
| Logistic Regression | 0.95 | 0.94 | 0.011 | 0.026 | 0.503 | 0.511 |
| KNN | 0.95 | 0.93 | 0.19 | 0.02 | 0.59 | 0.50 |
| Decision Trees | 0.95 | 0.94 | 0.09 | 0.01 | 0.54 | 0.50 |

Since all the models came with good accuracy score and bad recall scores, I researched some more regarding my steps forward. I came across the flow chart below that showed me how I should choose my metric for my current data set.



**Figure 7**

Looking at the Figure above, since my model is trying to predict class labels, and since both class are important. According to the figure above, it suggests that I used Accuracy, but I am not sure if that is the correct metric to choose as Accuracy will be a representation of predicting Stroke (0) – No stroke patients, as it is the majority class.

# Next process

Moving forward I will be measuring the models against ROC\_AUC score and might use smote as it did deliver good recall scores from KNN and Decision trees of about 84% on the testing data set. I will have to complete all the models and figure out which model accurately represents my data best. Also, once the best model has been found, feature importance can be used to understand which variables have the most impact on the model.