Stroke Predication

Note: After exploration, I found this data set is strongly imbalanced, this data set has 4861 records with target=0 (no stroke), but only 249 (less than 5%) records with target=1 (stroke).

There are six goals/needs in this project. So far, I achieved four from them which are:

- Find the factor has the most impact on stroke
- Find the most age had a stroke
- Does high blood pressure have an impact on stroke
- Use classification models to predict whether a given person had a stroke or not

1- Find the factor has the most impact on stroke:

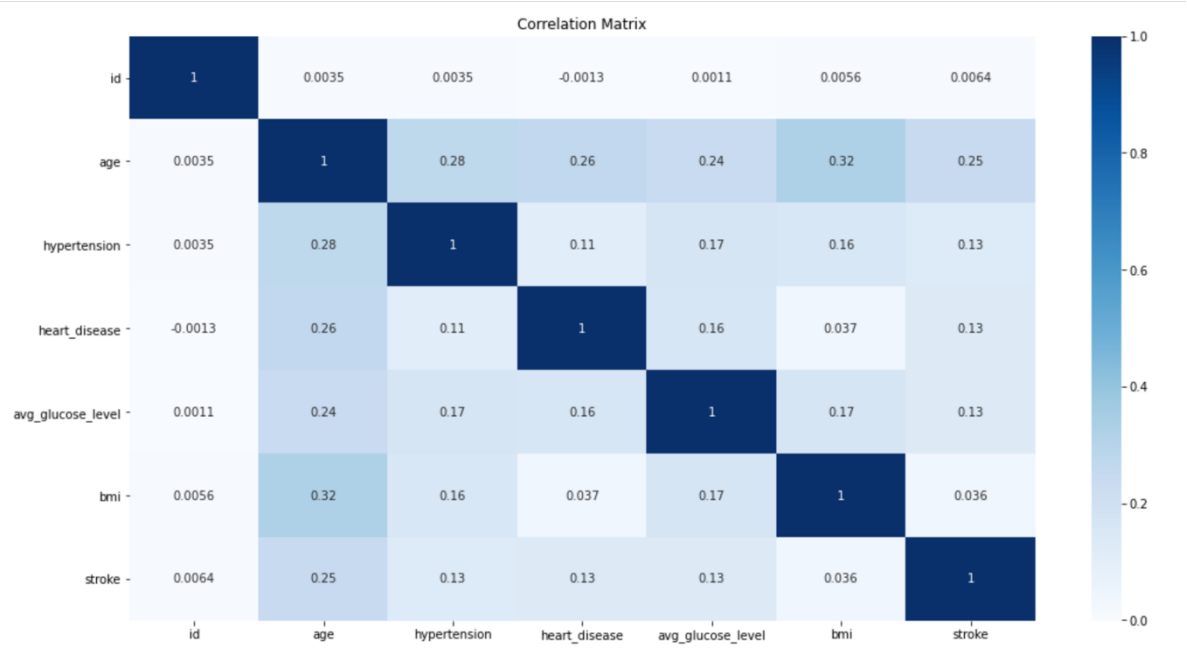

Figure1: Correlation Matrix

To start exploring this goal, I used a heatmap from the Seaborn library to find between features (gender, age, various diseases, and smoking status) and the target (Stroke).

The **Figure1** depicts that the heatmap makes a visual representation of the matrix. Each square of the heatmap represents a cell, the color of the cell changes according to its value.

From the **Figure1**, we can see that **age has the most correlation with the target**. Also, other variables have positive correlation values with the target except "gender", so it can be removed.

2- Find the most age had a stroke:

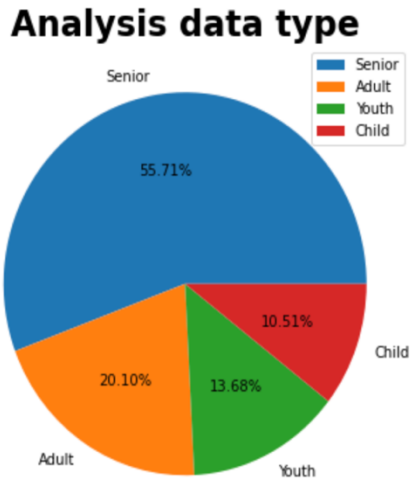

Figure2: Age Type

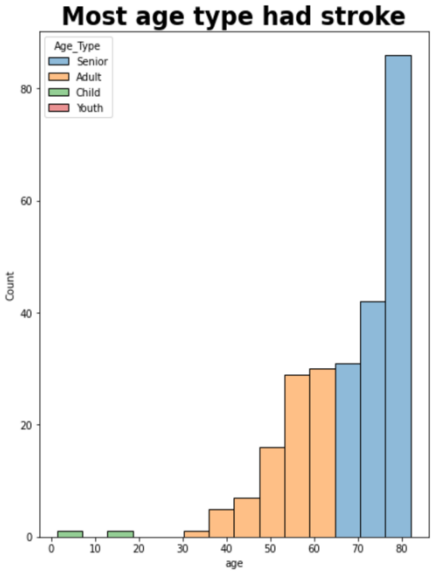

Figure3: Most Age Type had stroke

To start exploring this goal, first I categorized the age into 4 types (Children who are less than or equal to 14, Youth who are between 15 and 24, Adults who are between 25 and 64, and seniors who are more than 64). Then, I used the Histplot from the Seaborn library to find which the most category had a stroke.

Figure2 depicts the different age types in this dataset. **Figure3** describes the most age type had a stroke, in which X-axis represents the age in number and Y-axis represents the count of patients who had a stroke. The age types are classified by colors (Blue: Seniors, Orange: Adults, Green: Children, Red: Youth).

Figure3 shows that **people with age more than 64 are the most people who suffer a stroke**. While people with age type “Children” and “Youth” rarely have strokes compared to other categories.

3- Does high blood pressure have an impact on stroke:

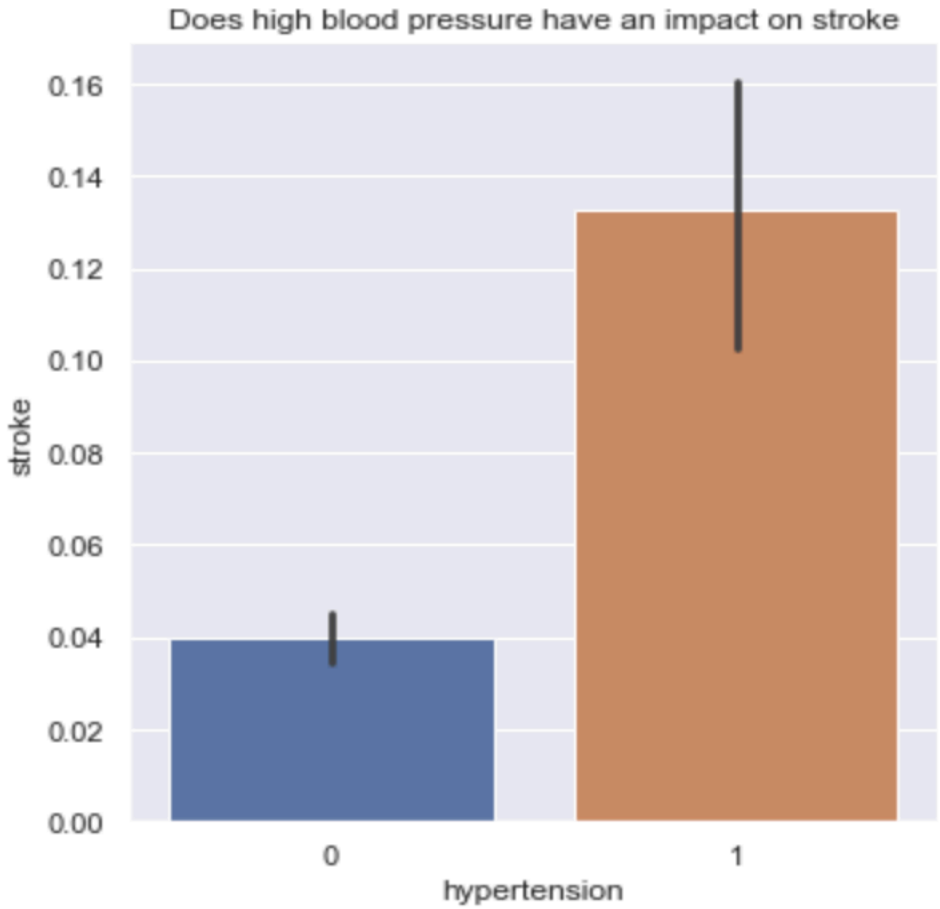

Figure4: high blood pressure has an impact on stroke

To start exploring this goal, I used the Barplot from the Seaborn library to find if high blood pressure has an impact on stroke or not.

Figure4 describes the effect of high blood pressure on stroke, in which the X-axis represents hypertension (0 if the patient doesn't have hypertension, 1 if the patient has hypertension), and Y-axis represents the mean of patients who had a stroke. Hypertension are classified by colors (Blue: Doesn't has hypertension, Orange has hypertension).

We can see from **Figure4** that people having more hypertension are more prone to stroke. That means the **main risk factor for stroke is high blood pressure**.

4- Use classification models to predict whether a given person had a stroke or not:

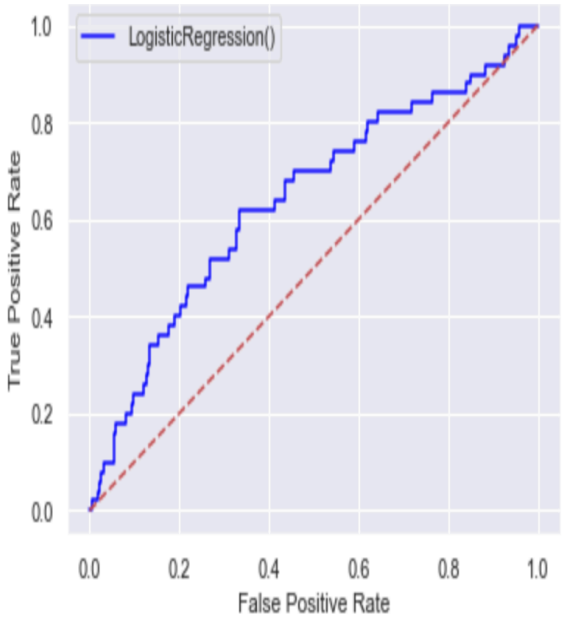

Figure5: LogisticRegression

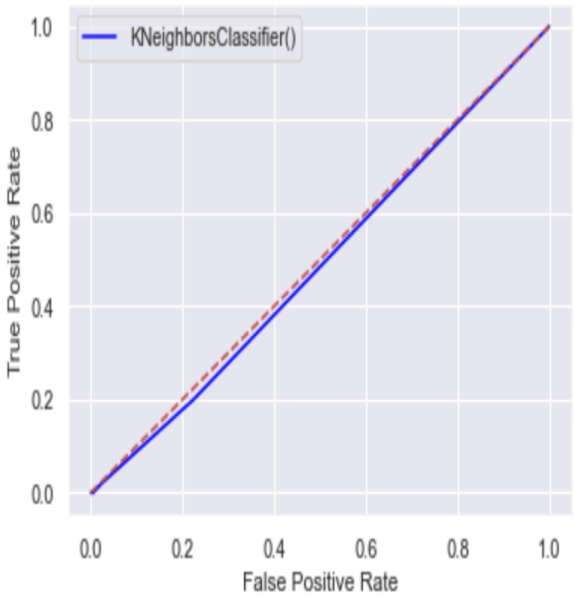

Figure6: KNeighborsClassifier

Since the data is imbalanced it could achieve high accuracy, if the classifier always returns 0 (4861/5110 = 95.13%). We can see that accuracy is not a useful metric in the context of strongly imbalanced data. So, I focused on AUC instead of accuracy.

The **Figure5** and **Figure6** depict the ROC curve of each model (blue) with True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on the y-axis and FPR is on the x-axis. An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability.

Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. So, the higher the AUC, the better the model is at distinguishing between patients with the stroke or not. For Now, **Logistic regression has an AUC score of 64%, which is good compared to KNeighbors Classifier which has an AUC score of 49%.** I should try another classification model, and I should tune the hyperparameters with the help of GridSearch to get the best model which has the highest AUC score.