Capstone 2: Stroke Data Report

Executive Summary

In this project, I worked with a stroke prediction dataset from Kaggle to implement a machine learning model that can predict whether a patient is likely to have a stroke based on various attributes such as: age, presence of disease, BMI, average glucose level, smoking status, etc. The dataset was tested against 7 different machine learning algorithms, and GridSearchCV was implemented to tune the models and find the best hyperparameters. The RandomForestClassifier turned out to be the best predictive model with an accuracy of 96.01% and AUC of 0.742.

Background

In the field of healthcare, much emphasis is placed on secondary and tertiary prevention, where an incident has already occurred or the patient has already been diagnosed and treatment is focused on preventing the disease from getting worse and reducing its symptoms. But a better solution would be to make primary prevention a priority in which we prevent the disease or incident from occurring in the first place.

According to the World Health Organization (WHO), stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths. Considering the cost of hospitalization, medication, and rehabilitation treatments, it is very expensive to treat a stroke patient. Life after a stroke can also be profoundly debilitating--with loss of mobility and cognitive function--oftentimes leading to an overall lower quality of life and depression. If the focus could be shifted into preventing a patient from getting a stroke rather than treating the patient after they've had a stroke, we could greatly improve a patient's quality of life and reduce treatment costs. A cost effective solution would be to identify risk factors that can increase the chances of a patient getting a stroke so that lifestyle modifications and prophylactic medication can be incorporated as primary prevention.

Data Cleaning and Wrangling

The dataset contains 12 attributes:

1. id: unique identifier
2. gender: classified as "Male", "Female", or "Other"
3. age: age of the patient
4. hypertension: 0 for absence of hypertension; 1 for presence of hypertension
5. heart_disease: 0 for absence of heart disease, 1 for presence of heart disease
6. ever_married: classified as "No" or "Yes"
7. work_type: classified as "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8. Residence_type: classified as "Rural" or "Urban"
9. avg_glucose_level: average glucose level in blood

10. bmi: body mass index
11. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"
    a. "Unknown" in smoking_status means that the information is unavailable for this patient
12. stroke: 1 if the patient had a stroke or 0 if not

The original data set had over 5,000 entries, but after dropping missing data, there were a total of 4,909 entries. It is noted that the data size is considered to be on the smaller scale, but still workable as it can be readily accessible and would provide valuable insight towards preventing a patient from getting a stroke. As the raw data was fairly clean, no further clean up or wrangling was required aside from dropping the missing data from the BMI column. Now the data is ready to be analyzed.

Exploratory Data Analysis

The variables were separated by numerical and categorical features. The stroke column was identified as the target variable.

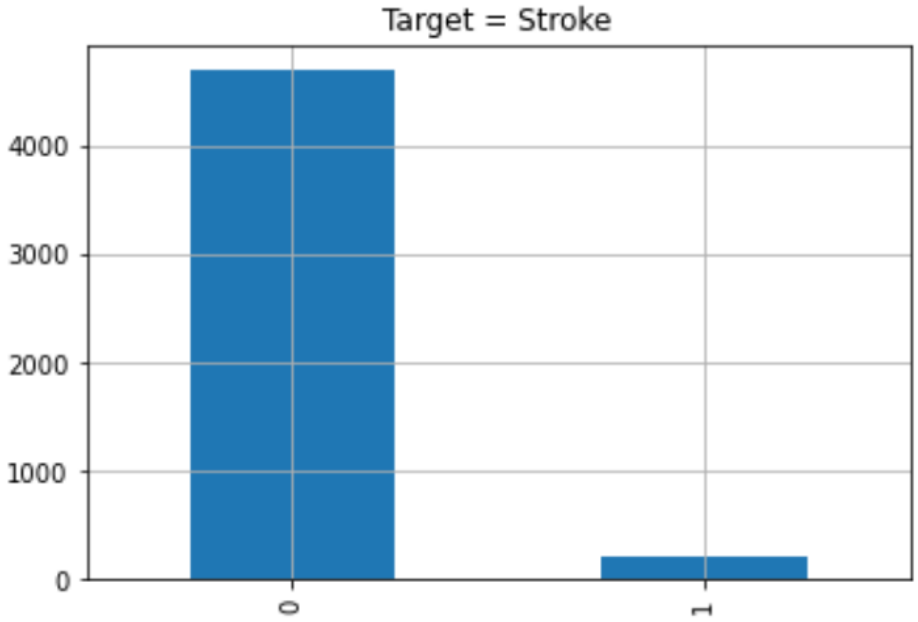
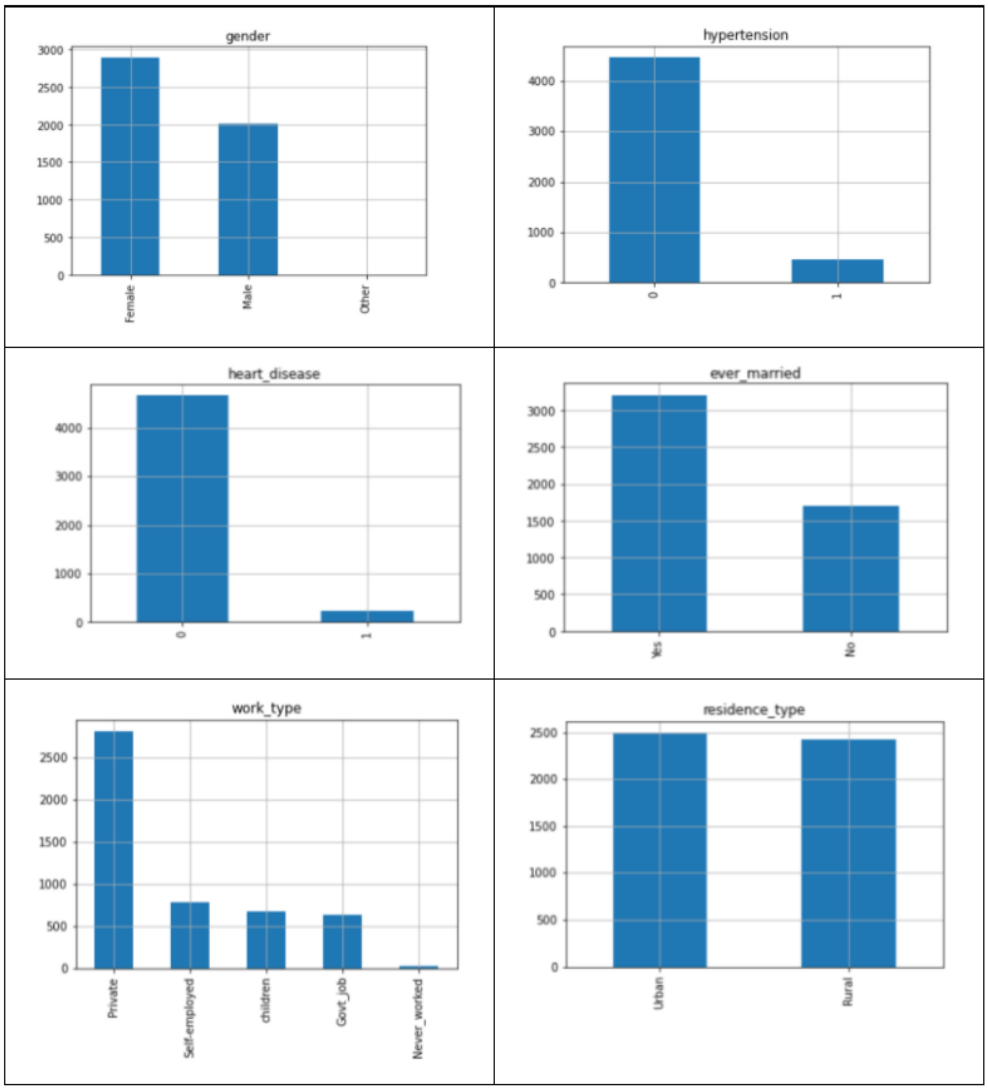


Figure 1: Target distribution plot. There are a lot more patients that did not suffer from a stroke (4700) than patients that did have a stroke (209).



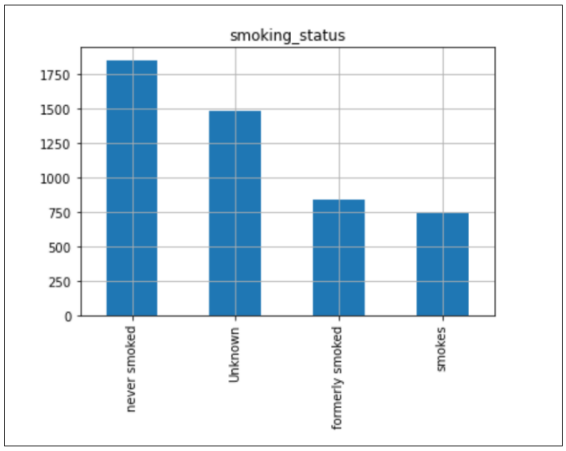


Figure 2: From the distribution of the categorical features, it can be inferred that the dataset contains more women patients than men. The majority of patients do not have hypertension or heart disease, are single, work in a private setting, and do not smoke or their status is unavailable. There was a pretty even distribution between residents that live in an urban and rural area.

The distribution of the categorical features shows the dataset seems to represent mainly healthy patients that do not have any heart complications.

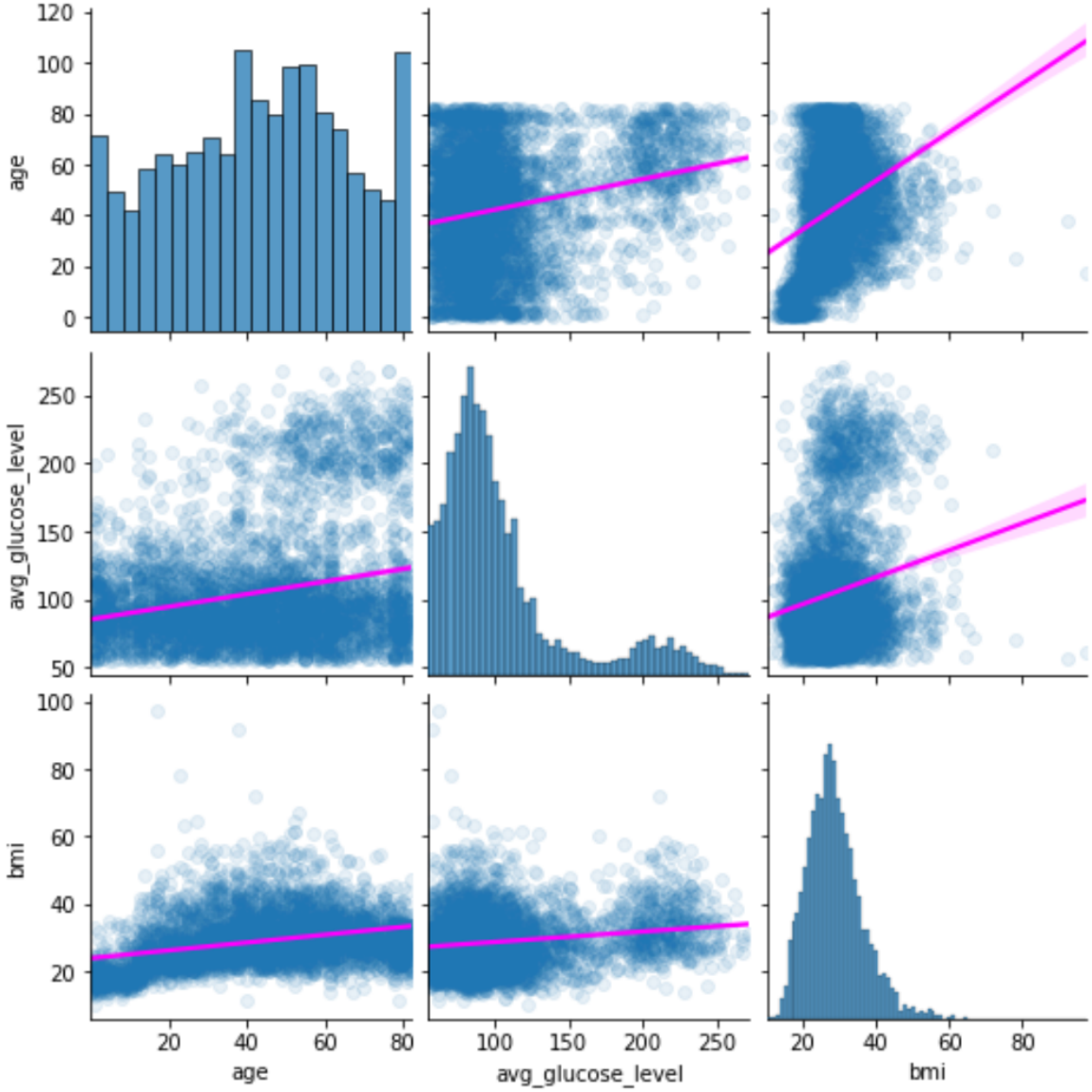


Figure 3: A pairwise scatter plot between the numerical features. It seems as age increases, so does the average glucose level and BMI.

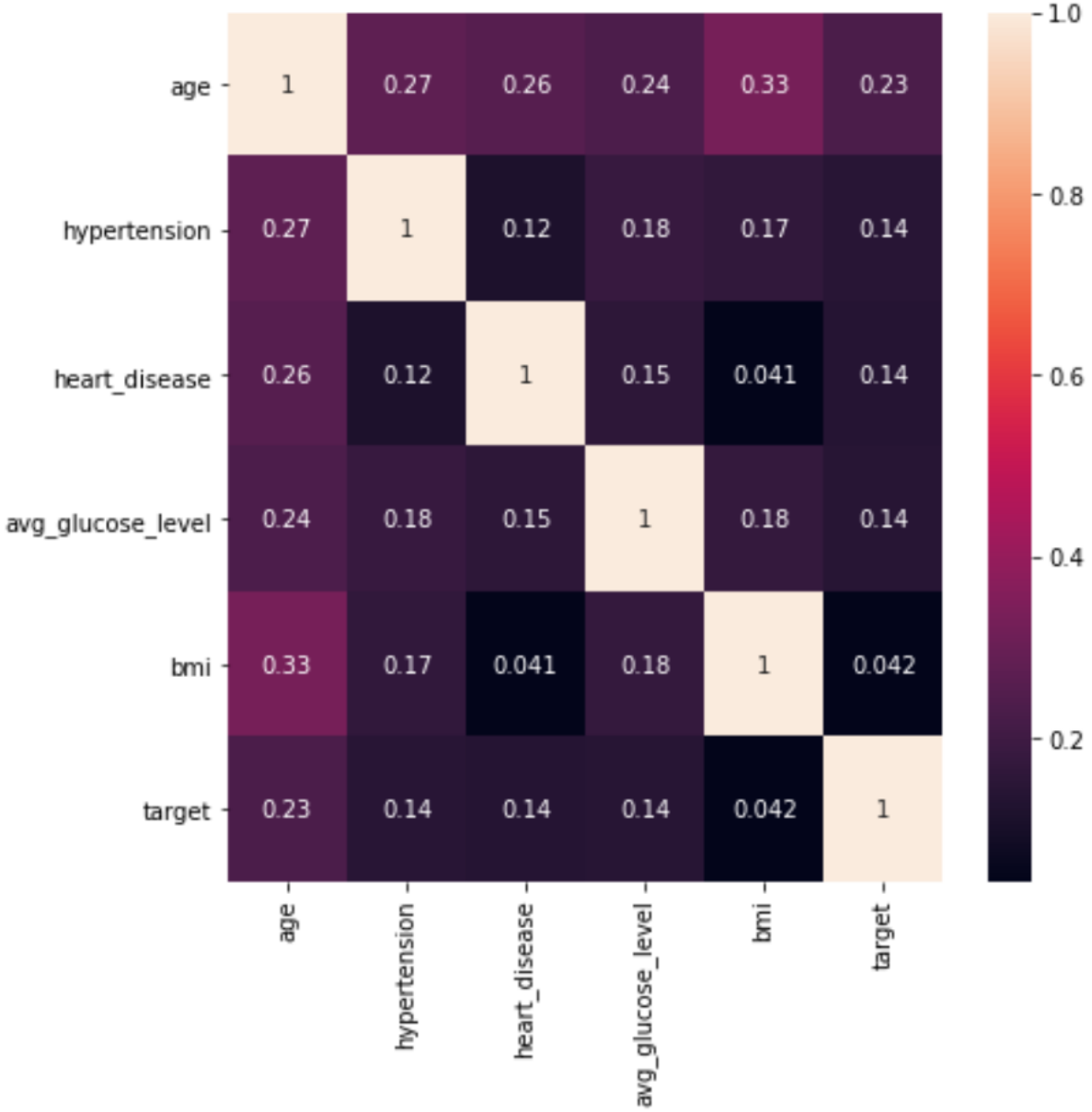


Figure 4: A heat map to analyze the correlation between the various attributes. It can be noted that age has the highest positive correlation (0.23) with our target variable (stroke).

The pairwise plot and heat map shows that many features were weakly correlated but none were so strongly correlated as needed to be removed.

Model Evaluation

GridSearchCV was used to select the best hyperparameters and fit on the models.

Model	Accuracy(%)
GaussianNB	19.35
BernoulliNB	94.06
LogisticRegression	96.04
RandomForestClassifer	96.01
DecisionTreeClassifier	92.43
KNeighborsClassifier	96.04
SVC	96.01

Table 1: The accuracy of the models that were tested. LogisticRegression, RandomForestClassifier, KNeighborsClassifier, and SVC all had accuracies over 95%.

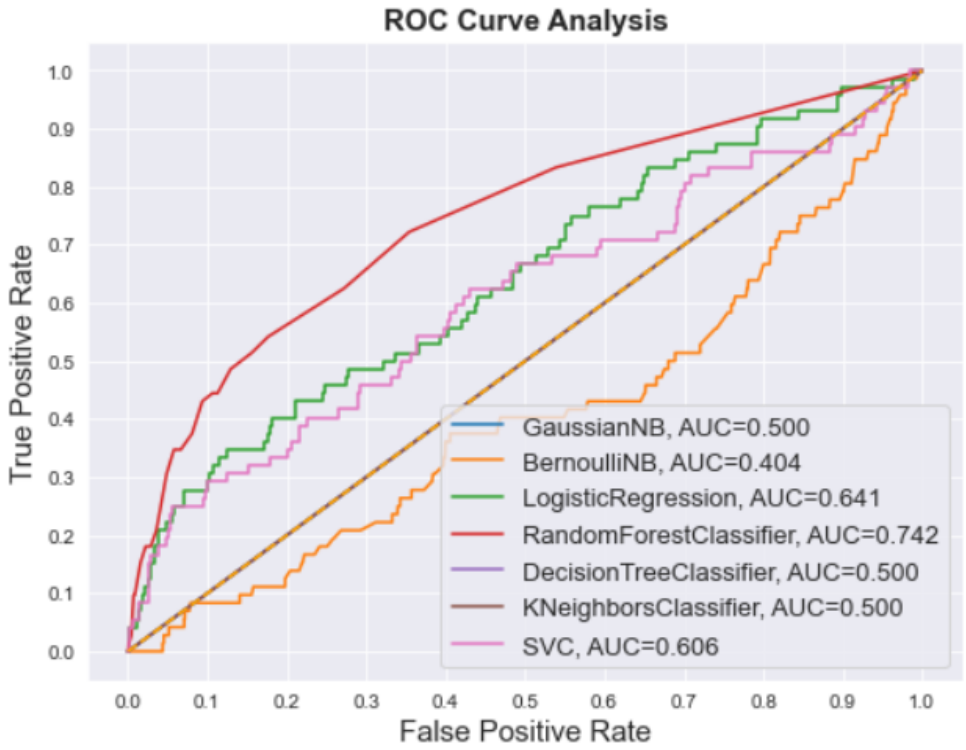


Figure 5: ROC curve plotting the false positive rate against true positive rate. RandomForestClassifier shows the highest AUC of 0.742 while LogisticRegression, KNeighborsClassifier, and SVC all had significantly lower AUCs.

Although the LogisticRegression, RandomForestClassifier, KNeighborsClassifier, and SVC all showed accuracies over 95%, the ROC curve shows that RandomForestClassifier had the highest AUC from all the models tested.

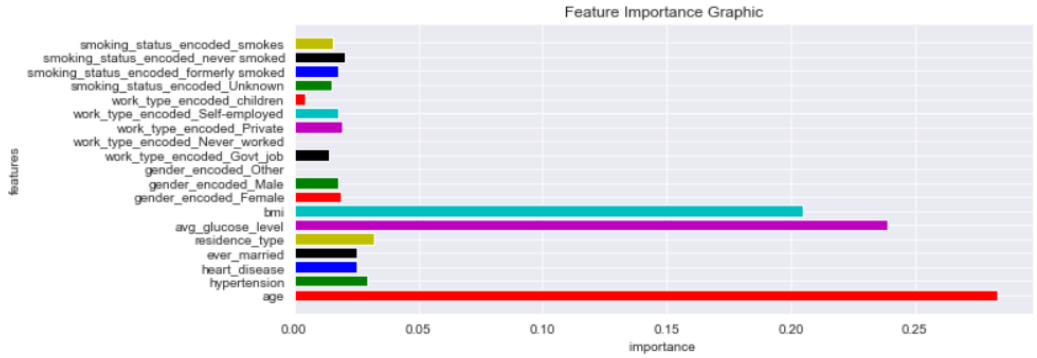


Figure 6: Horizontal bar graph depicting the importance of the various attributes. Avg_glucose_level, age, and bmi are the top three most important features.

The heatmap (Figure 4) previously showed that age is correlated to stroke, and this graph further confirms it as age, avg_glucose_level, and bmi are the top three important features.

Takeaways

The best model for this stroke prediction dataset with the highest accuracy and AUC of 96.01% and 0.742, respectively, is the RandomForestClassifier model (Table 1, Figure 5). This model was used to find the features that were most important to the target (stroke) variable. As depicted in Figure 6, age, avg_glucose_level, and bmi are the three top most important attributes.

Using attributes that are readily accessible, data for this model can be easily collected at doctors' visits and quickly return results without requiring invasive testing or waiting for lab work turnaround time. This model would be useful for health care professionals to efficiently identify patients that are at high risk of getting a stroke and determine which variable contributes to the high risk. This information can be used to counsel patients on how to lower their risk, either through lifestyle modifications or with preventative medication. By identifying risk factors early on, primary prevention could greatly reduce the cost of treatment for a stroke. This predictive model would be highly useful for focusing on primary prevention of a stroke and therefore contribute to providing optimal patient care by reducing treatment costs and pill burden, ultimately improving a patient's quality of life.

Going Forward

When coming in for a routine check up with your primary care physician, if the model predicts that you will get a stroke based on your attributes (older age, uncontrolled blood glucose, high BMI, etc.), an intervention can be placed to prevent your condition from getting worse. Age is not a factor that can be controlled, but lifestyle modifications and prophylactic

medicine can help with high blood sugar and a high BMI. Together with the patient, the doctor can discuss possible solutions for a healthier lifestyle. This can include moderating the patient's diet and encouraging exercise so they can reduce their weight to a normal range. If diet and exercise alone are not enough, the patient can be further assessed to see if they need to be started on statin or antiplatelet medication to further reduce their risk. If the patient is already on these medications and does not see any improvements in their health, it should be checked if the patient is on the appropriate medication regimen (and correct dosage) and whether the patient is in compliance with their medications.

Many steps can be taken to prevent a stroke from occurring, and by identifying these risk factors early on to make small modifications, a costly stroke can be prevented. In the US, the average cost of hospitalization for a stroke is around $20,000. On the other hand, statins and antiplatelets such as aspirin are covered by most insurances and cost a fraction the price of a hospital visit. Even in the event of a false positive, advising a patient to strive for a better blood sugar level or BMI would not bring the patient any harm. On the contrary, life after a stroke can be very difficult to adjust to. If the patient does not fix their lifestyle habits, they are at risk for getting a second stroke. By implementing these interventions early on, patients are equipped with better knowledge to be healthier and live more fulfilling lives.

To improve the quality of this dataset and the machine learning models, more data from patients that did suffer from a stroke should be collected. This would improve the distribution of the data as the original dataset had an overwhelming amount of healthy patients that did not have a stroke. Since the size of the dataset is also small, more data will improve the integrity of the data. The demographics of the patient should also be included to get a better picture of what type of population the data best represents.