Stroke Prediction Project - CMSE 492

Sanjay Subramanian

Background and Motivation

Repository Link: https://github.com/sanjay7303/cmse492_project/

The following project is to identify what are the main features that affect whether a patient has a stroke or not. The following features are hypertension, bmi, work_type, smoking_status and many more. Moreover, the idea is to use machine learning to predict which model for what high quality features can produce an accurate prediction. The following dataset was retrieved from kaggle and by using machine learning it offers us several benefits by allowing us to handle complex relationships between the many features as listed previously and also allow us to identify the non-linear relationships between them as well. For this specific project, the data that I am using is considered as static data as it is only from a specific time frame but as the data changes or becomes dynamic the machine learning models can run continuously to give us constant predictions. This was one of the main motivators for me to take on this project.

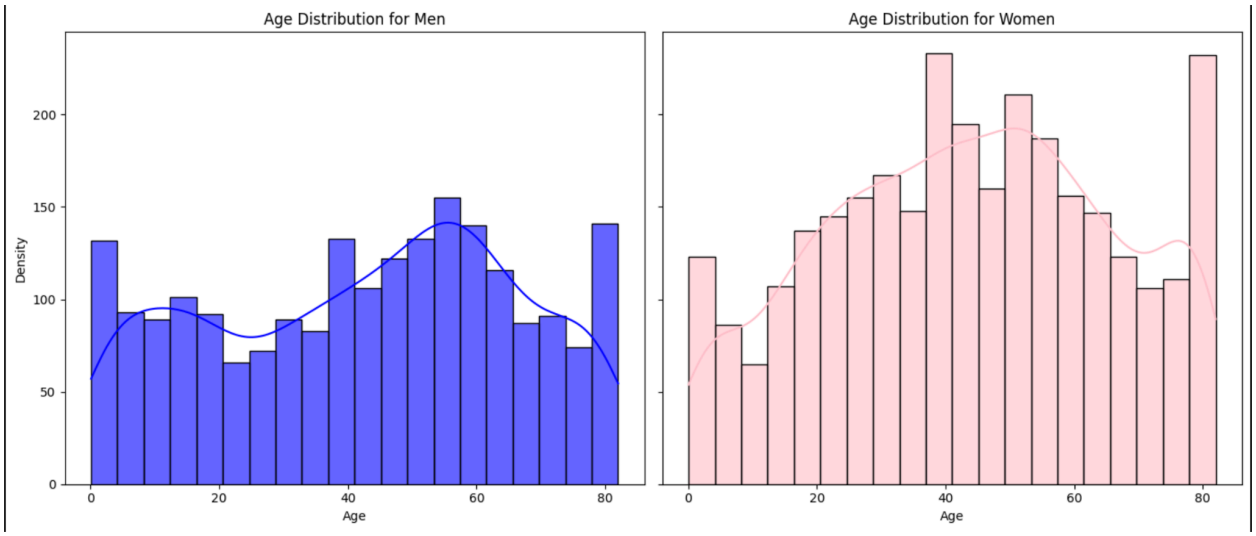
Metrics, ML Task and Objective

The Machine Learning task for this following project is to determine whether a patient is at risk of experiencing a stroke based upon the following features in the dataset. Most of those features are lifestyle based and medical history information as well. Furthermore, the objective of this project is to classify a patient's risks of obtaining a stroke or not, which can be classified as a positive class and a negative class. Since this can be considered as a binary classification, using Machine Learning is very much appropriate.

The success of the model is determined by the test accuracy score as well as the f1 score, precision and recall. With regards to the test accuracy, a score of 0.95 and above would be considered as an accurate machine learning model. Moreover, a higher recall is something to consider ensuring the cases of patients that have a stroke are not missed. I will also be taking a look at the confusion matrices of all the models and see whether the true predictions are found for most of the patients. Lastly, another metric I decided to use is the AUC score. Since this is a classification prediction project using that metric would help us differentiate the performances of the different models. I decided to say if the AUC score is greater than 0.80 then the model is considered excellent.

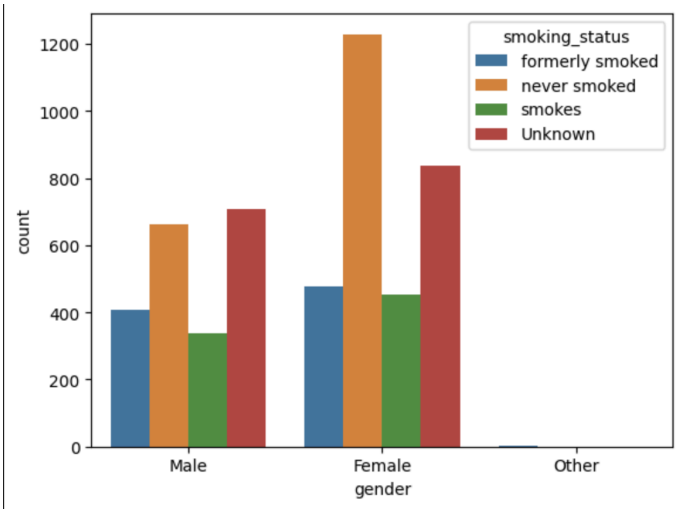
Initial and Exploratory Data Analysis

For the initial data analysis, I thought it was imperative to take a look at the age distribution between the male and female patients from the data. To do so, I decided to plot an histogram for each. The results are shown below.



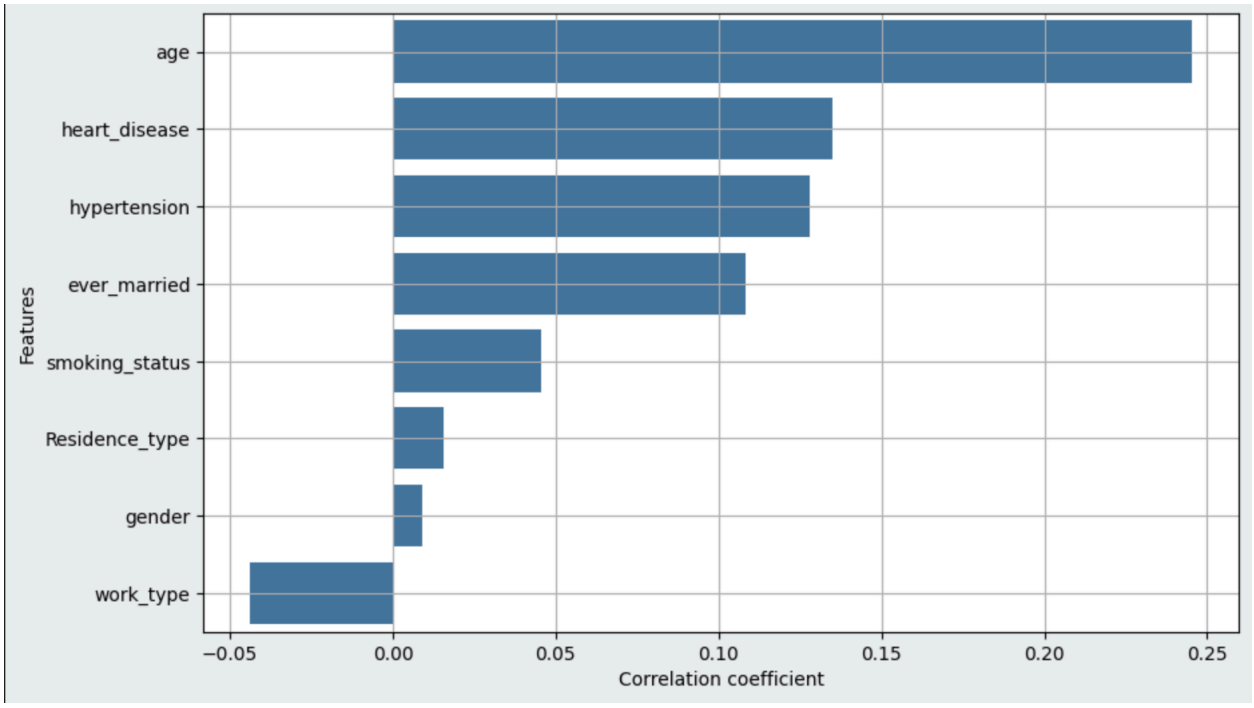
From the above plots we can see that there is a higher count of females in this dataset with a large proportion of females in the age range of 80. Whereas in the male age distribution graph there is a more even spread. For men, the age distribution seems to be very uniform with peaks present in the age category of 75-80 as well as from 0-10. With regards to the age distribution for women, there is a stronger peak in the middle age range from 40-60.

Next, I wanted to analyze the smoking status of all the patients and the difference between males and females in that category. The results are presented in the illustration below.



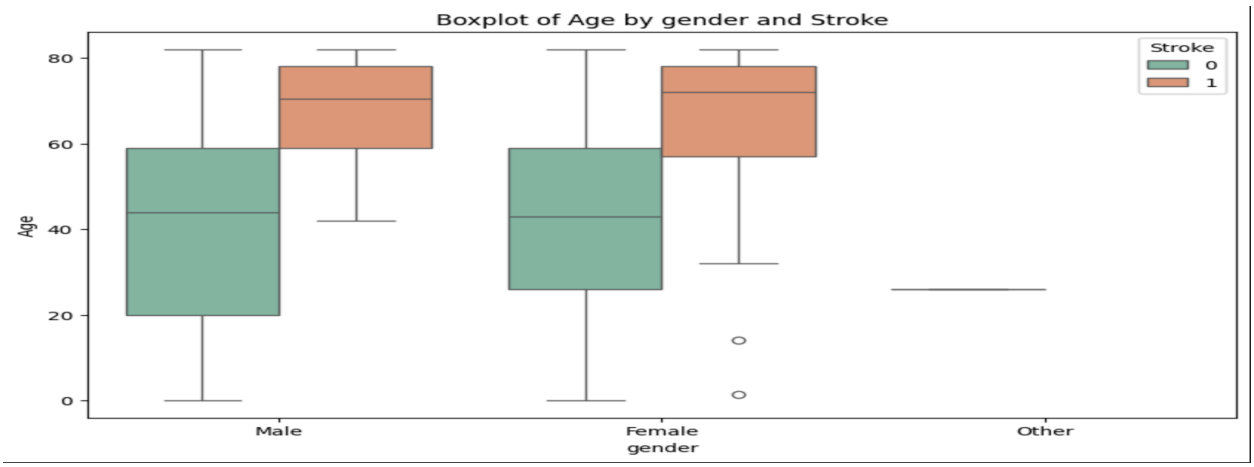
From the above plot, you can see that more females have never smoked compared to males and the other categories are more or less the same.

Next, I wanted to see the correlation of all the categorical variables against the target variable which is stroke. Reason for this is to deduce which variable is the most important when we predict the model. The results are illustrated below.

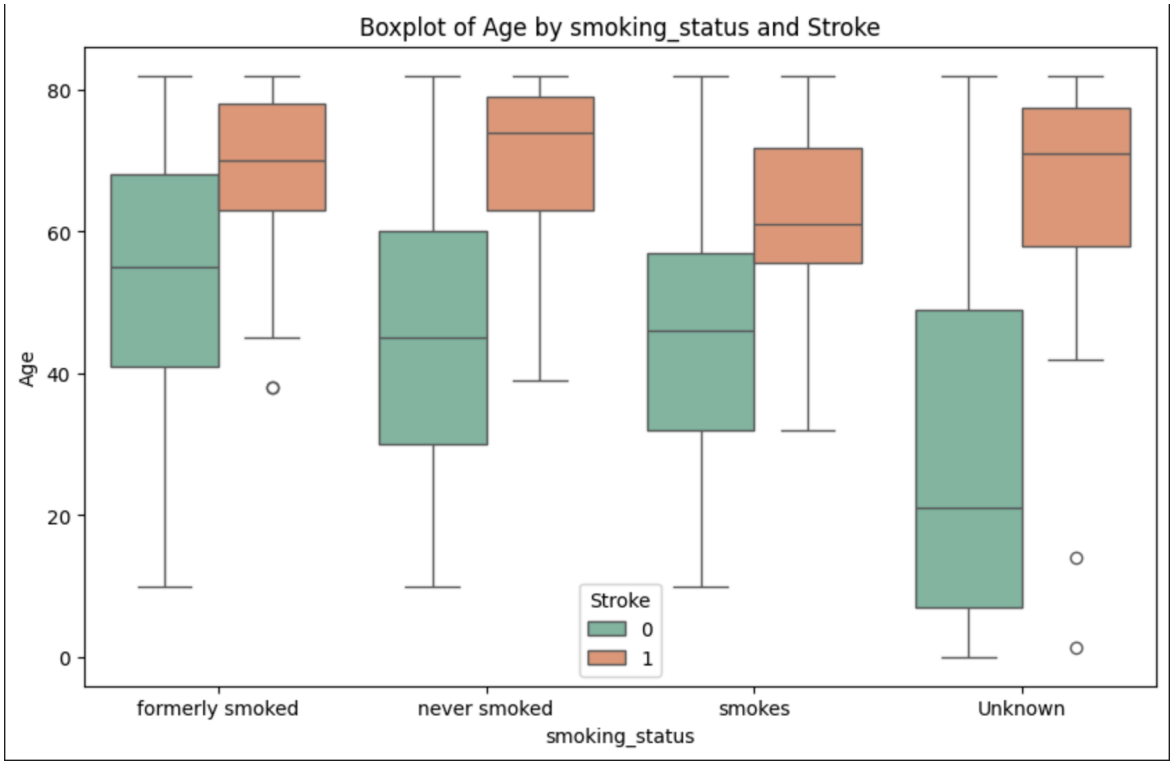


From the illustration above we can see that all of the features except work_type have a positive correlation but from that age is the feature that is highly correlated. We can then deduce that age is the most important categorical feature from the dataset.

Using that I then decided to create a boxplot and visualize the distribution in relation to stroke and gender. The results are illustrated below.



From the plot above you can see that from the male and female categories of patients, the plots indicate that patients with a stroke are of a higher age range compared to the patients that do not have a stroke. The same plot was created with the feature smoking status.



From this plot as well you can see the patients with a stroke are of a higher age and the distribution is consistent with the previous boxplot as well.

Models and Training Methodology

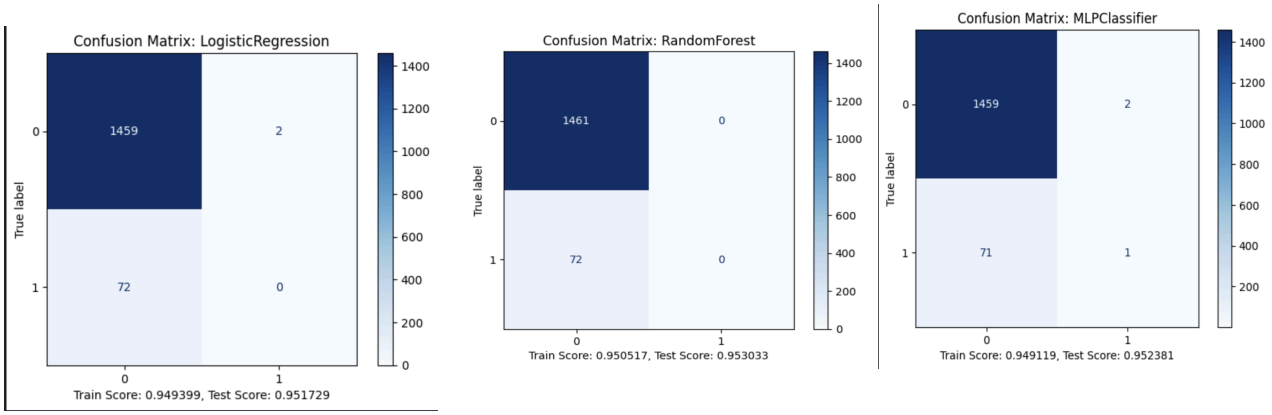
For this project, I chose a variety of models to best interpret the prediction with this given dataset. As for the baseline model, I chose it to be a logistic regression. The reason being it is well versed with binary classification as it is with the data that I am dealing with in this project, with predicting whether a patient has a stroke (1) or not (0). Moreover, I chose it due to its ease of implementation as it does not require a high amount of hyper parameter tuning. Next, the other models I chose for comparison are Decision Tree, Random Forest, MLP classifier, Support Vector Classifier and Gradient Boosting. Reasons for choosing these models are because they are more robust compared to Logistic Regression and they allow for some variability when it comes to interpreting non linear relationships within the data.

Prior to implementing the model and running the model, it is imperative to wrangle the data and make sure it is in the format to produce an accurate prediction. Initially when working with the data I was able to find that there were some missing values in the BMI column. For better model accuracy it is very important to replace those missing values rather than simply removing them. So, I did that by using the Simple Imputer library in SK learn and replace the values with the mean of the remaining BMI values in the existing column. Next, is dealing with the categorical variables. Categorical variables like ever_married, work_type, Residence_type and smoking_status in the dataset are not numeric which will not allow us to feed it into our models. To combat that, I decided to use One Hot Encoding which is a strategy to convert the categorical variables into numeric and allow the machine learning models to understand the data a lot better.

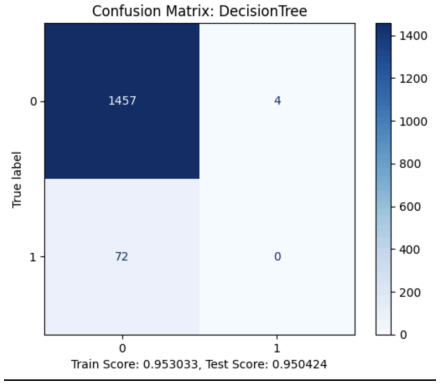
Next, when preparing my models for running I used train_test_split from the sklearn library to split the data into training and testing. I also removed the work_type column from the training and testing data as it had a negative correlation to the target variable which will most likely skew the accuracy of the results for predictions. Furthermore, I used cross validation and hyper parameter tuning for my models. Using the random_search_cv function in python, I was able to tune the parameters for all of the models. Since I had to perform cross validation across many models, I thought it would be used to go for random search cross validation as it would not take a computationally longer time compared to grid search cross validation.

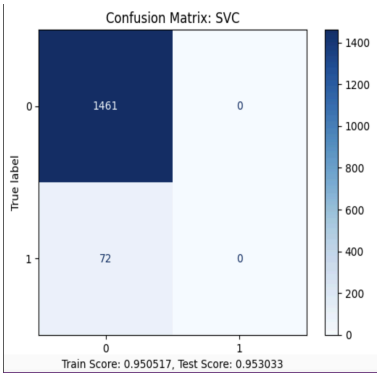
Results and Model Comparison

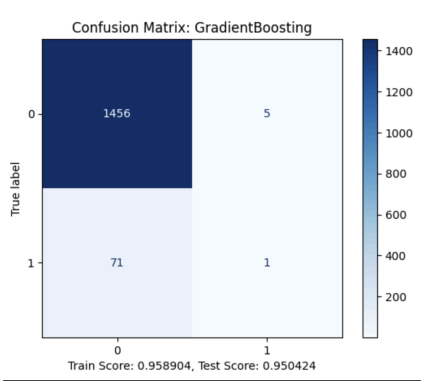
To visualize the model performance and accuracy, I decided to use the confusion matrix to see the true predictions and as well as the ROC curve. Additionally, I used the test accuracy scores to visualize them using a histogram. The results are illustrated below.

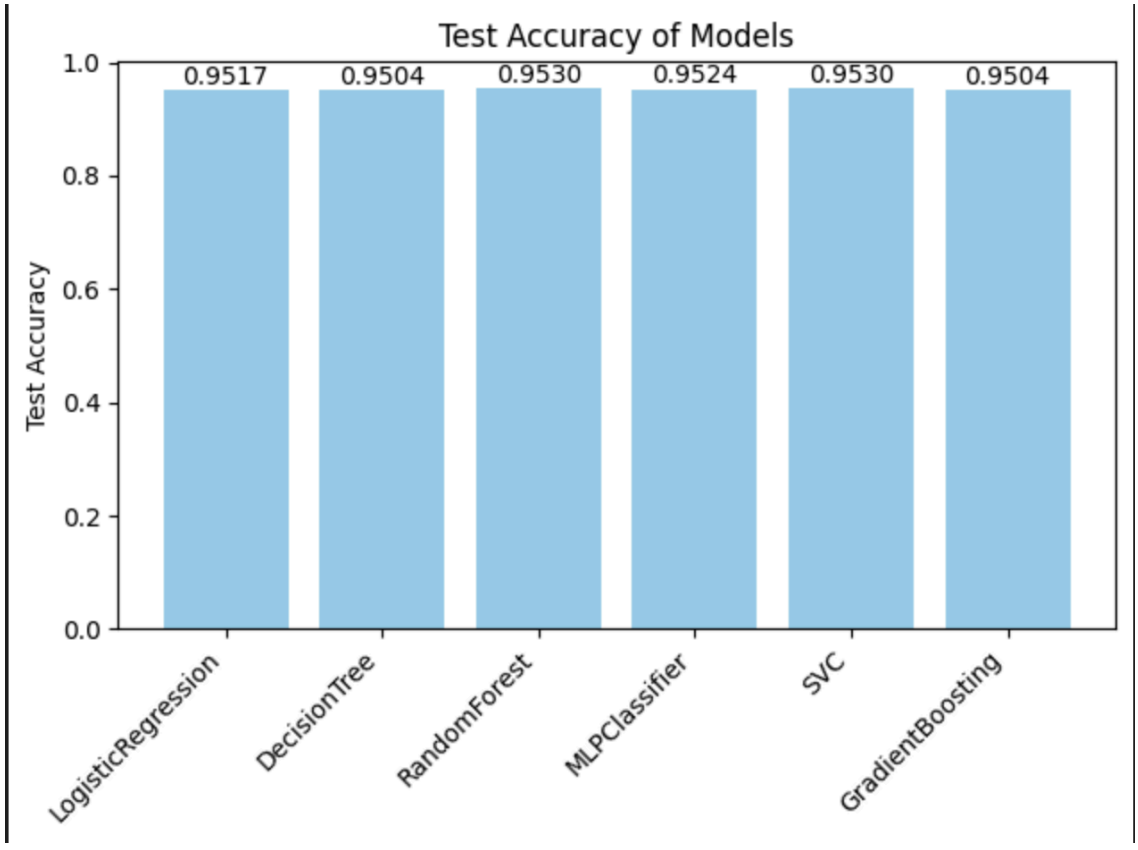


For the 3 confusion matrices above we can see that the random forest model has the highest test accuracy score of 0.953033 and Logistic regression has the lowest of 0.951729. But as per the metric decided earlier with the accuracy being greater than 0.95, all of the models satisfy that, so we can say that they are all highly accurate.

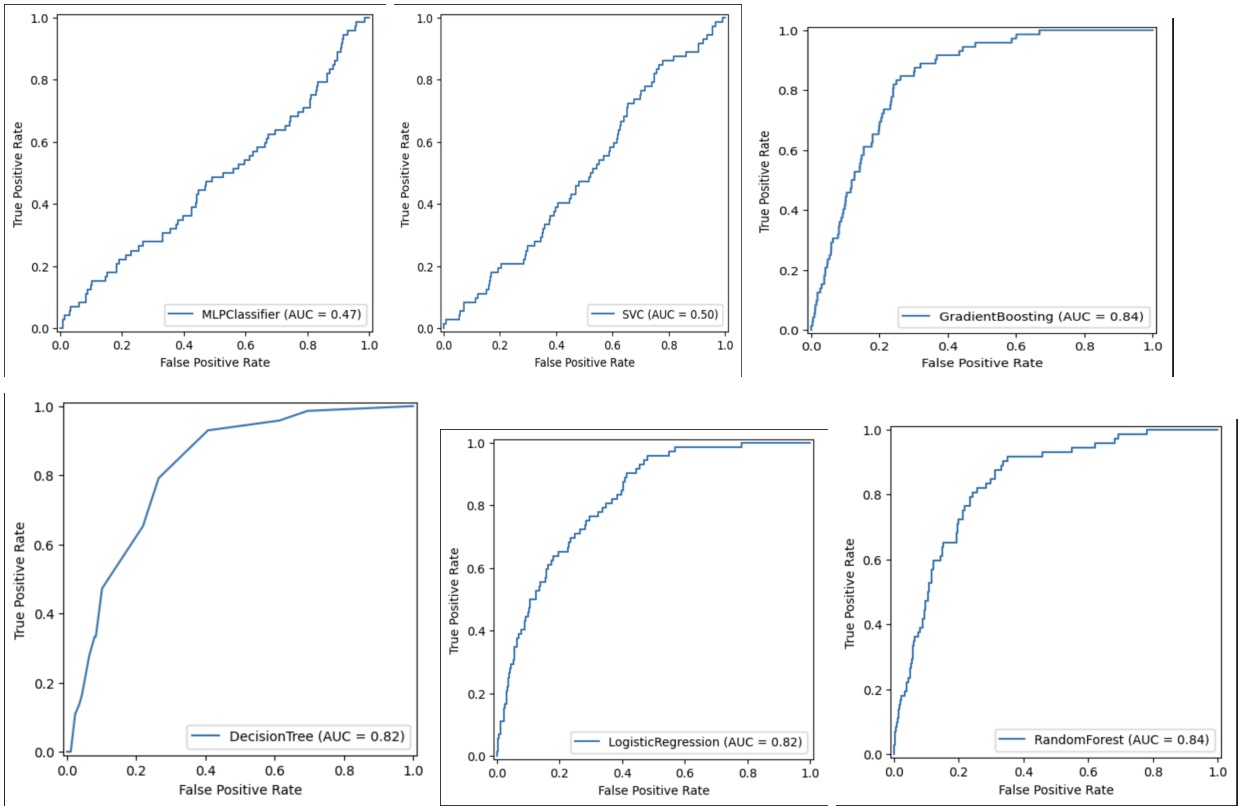








From the above illustrations we can see that almost all of the models have a high test accuracy with Random Forest and SVC being the highest with 0.9530. From this metric we see all the models are highly accurate, but for more conclusive evidence, I decided to see the ROC curve with the AUC score. The results are below.



From the following plots we can see that the random forest model has the highest AUC score. From both metrics decided we can see that it is outperforming the others. So, we can conclude that the Random forest model is the best. The reason for it being the best is because it is very robust and prevents overfitting unlike the other models. With regards to SVC having a high test accuracy and low AUC is because of the dataset being imbalanced with high amounts of data present of patients without a stroke compared to patients with a stroke. This results in the SVC model in just predicting the majority class of patients with a stroke and the AUC score and ROC curve reflects that completely. This can be said with the MLP classifier as well with it being not so confident in its predictions.

Conclusion

In conclusion, this project was a major eye opener with regards to model performance and trade offs. I was able to see what sort of planning and inference is needed in order to achieve the best performance for the very many models present. When doing this project I was able to use my metrics to great use and see which models satisfy both of those metrics to the best of its ability. Random Forest did that very well with it having the highest AUC score as well as test accuracy. Since this data is imbalanced, Random forest uses its ensemble approach by combining multiple decision tree models to reduce the imbalanced nature of the data. That was one of the major factors of why the model was the most effective.

With regards to limitations, the data was imbalanced allowing for the negative class to have a higher amount. Because of this, it allowed some other models like SVC to not be as accurate as it should be in other metrics. Moreover, the dataset in general was small with only around 5000 records which resulted in the data not being represented enough.

Lastly, for improvements, it would be wise to oversample the negative class to ensure the data is balanced for an accurate prediction. Using grid search cross validation would also be better due to its exhaustive approach in hyper parameter tuning. Additionally, using complex neural networks could present a better model prediction compared to the ones chosen and tested.

References

1. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data
2. https://scikit-learn.org/stable/supervised_learning.html
3. https://msu-cmse-courses.github.io/cmse492-FS24-jb/intro.html