Stroke Prediction

Leena Alotaibi | Classification Project

Goals

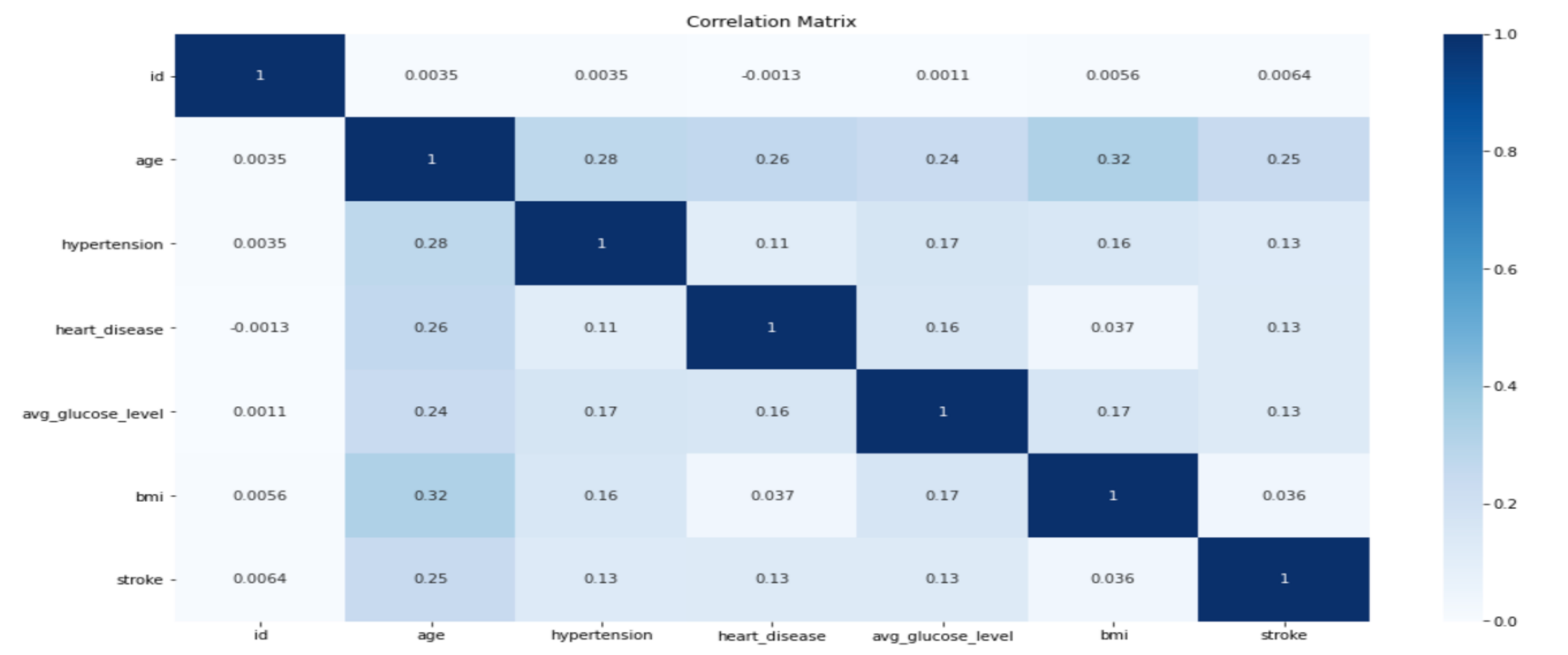
The purpose of this project was to use classification models to predict whether a given person had a stroke or not, and find the main factor for stroke to create a prediction system to predict the stroke in its early stages.

Data

This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Findings

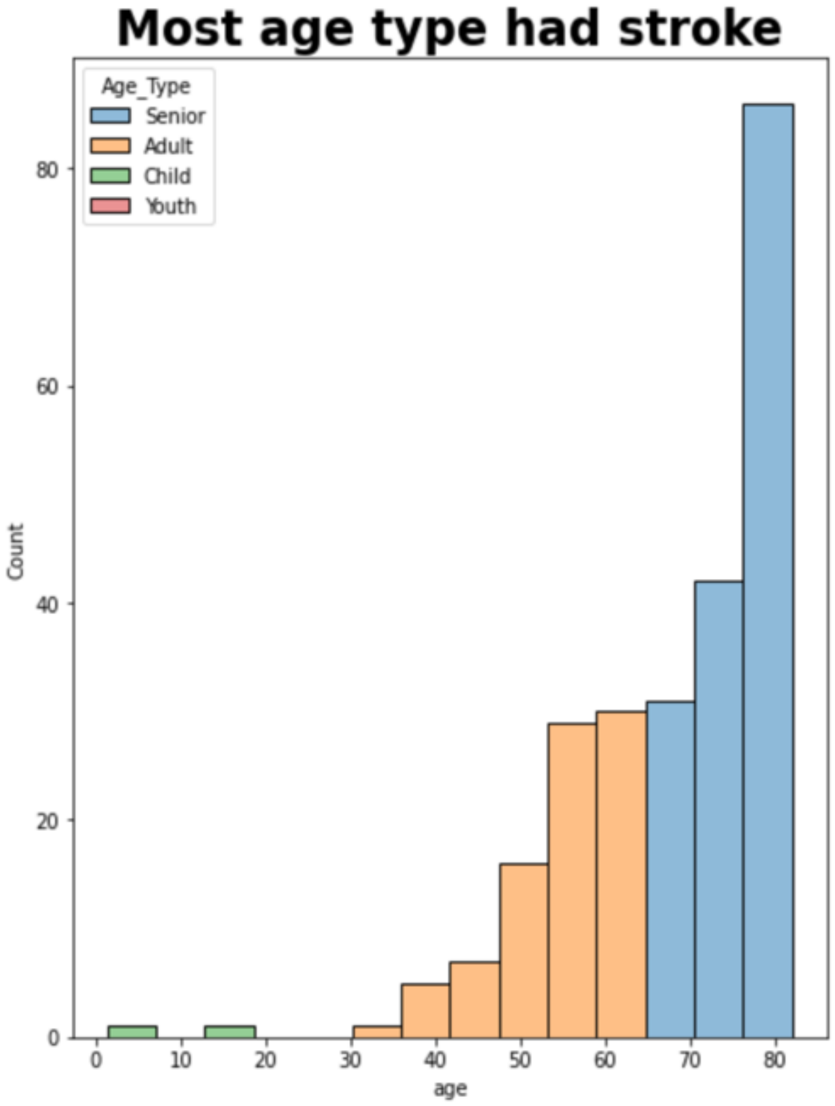
Age has the most correlation with the target. And other variables have positive correlation values with the target.

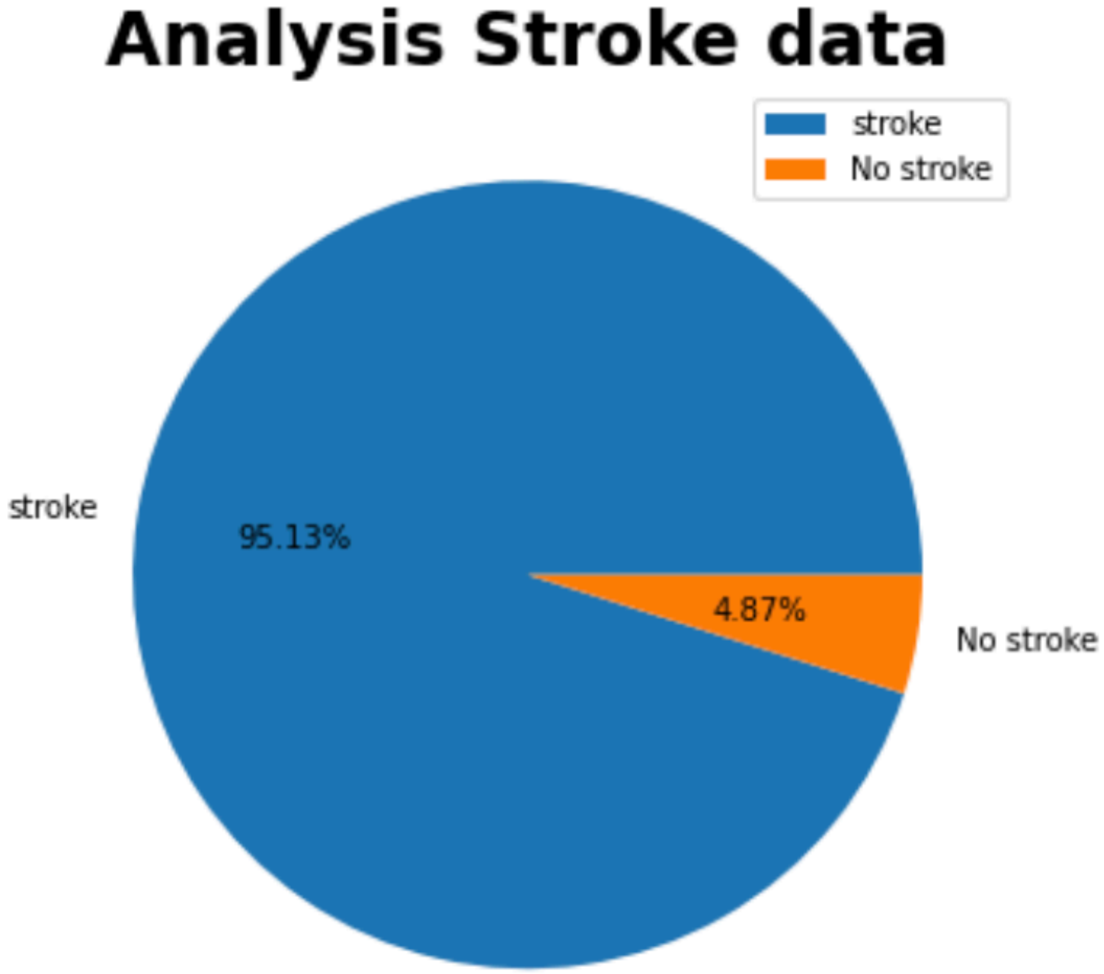


Correlation Matrix

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
id	1	0.0035	0.0035	-0.0013	0.0011	0.0056	0.0064
age	0.0035	1	0.28	0.26	0.24	0.32	0.25
hypertension	0.0035	0.28	1	0.11	0.17	0.16	0.13
heart_disease	-0.0013	0.26	0.11	1	0.16	0.037	0.13
avg_glucose_level	0.0011	0.24	0.17	0.16	1	0.17	0.13
bmi	0.0056	0.32	0.16	0.037	0.17	1	0.036
stroke	0.0064	0.25	0.13	0.13	0.13	0.036	1

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.





Imbalanced data

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).

I handled Imbalance data using SMOTE(resample)

Model

Model Evaluation and Selection

I used Logistic Regression, Decision Tree, K Nearest Neighbour, and Random Forest Classifier

I tuned the hyperparameters with the help of GridSearch to get the best model based on highest performance (AUC or Accuracy)

Before resample data VS. After resample data

Before resample data

Since the data is imbalanced it could achieve high accuracy, if the classifier always returns 0 (4861/5110 = 95.13%). I focused on Area Under the ROC Curve (AUC) instead of accuracy

Model that has the highest AUC score of is Random Forest Classifier

Model	AUC Score
Logistic Regression	68%
Decision Tree	56%
K Nearest Neighbou	64%
Random Forest Classifier	84%

Before resample data VS. After resample data

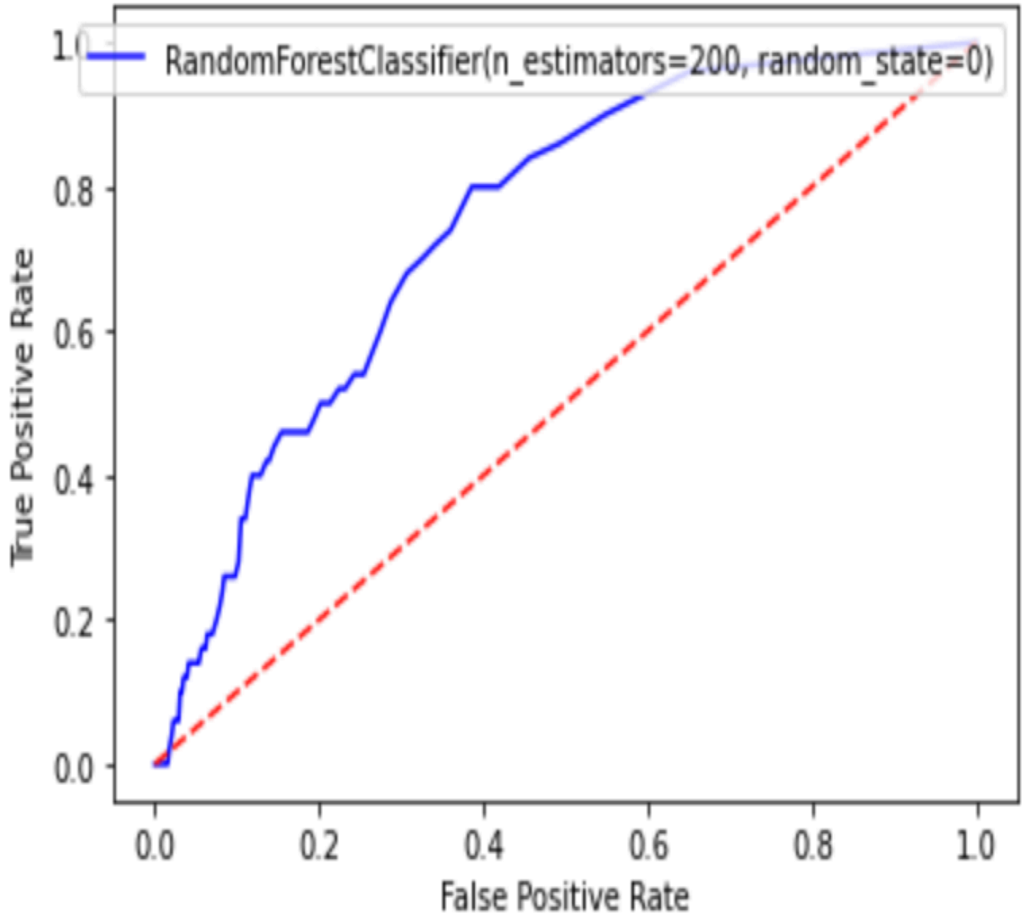
After resample data

In this case I focused on accuracy.

Model that has the highest accuracy score of is also Random Forest Classifier

Model	Accuracy Score
Logistic Regression	86%
Decision Tree	91%
K Nearest Neighbou	85%
Random Forest Classifier	95%

Use classification model to predict whether a given person had a stroke or not



Good measure of separability

Conclusion

- Data visualization and data analysis help in finding the factor that has the most impact on stroke and finding the most age had a stroke
- There is no strong correlation between features and target. Although each column had a positive correlation with the target variable, none of them was higher than 0.5.
- In unbalanced data we have AUC score of 84% in Random Forest, which is good
- After resample data we have the highest accuracy of 95% in Random Forest

Thank you