YALE UNIVERSITY COMPUTATIONAL METHODS FOR INFORMATICS (BIS634)

Final Project Report

Stroke Prediction and Analysis

Cao Zhiyuan

December 21, 2022

Partners:
Name: Cao Zhiyuan NedID: zc347

Contents

1 Introduction

A stroke is a medical condition in which the blood supply to a part of the brain is disrupted, leading to brain cell death and possible long-term disability or death. Stroke is a leading cause of death and disability worldwide, with high rates of morbidity and mortality. It is worth paying attention to stroke because it can have a significant impact on an individual's quality of life and can also have a significant economic burden on society.

In order to better understand and ultimately prevent or treat stroke, it is important to analyze data on stroke incidents and outcomes. Choosing a dataset for stroke analysis can help researchers and healthcare professionals gain insights into the risk factors, causes, and consequences of stroke, as well as identify potential interventions or treatments that may be effective in reducing the incidence and severity of stroke. By analyzing data on stroke, we can improve our understanding of this important public health issue and work towards finding solutions to reduce the burden of stroke on individuals and society.

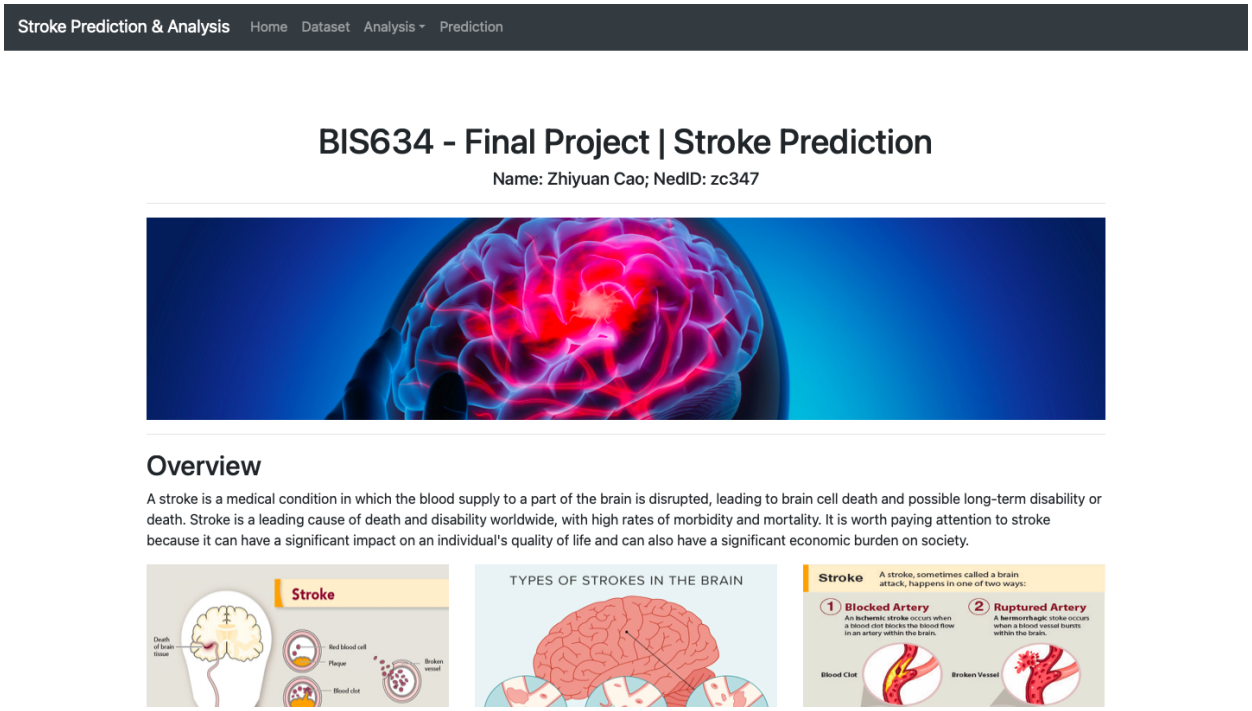


Figure 1: Webpage for Homepage

2 Dataset Description

2.1 About the Dataset

The dataset named "Stroke Prediction" is an open source dataset from Kaggle. It is under the healthcare-dataset-stroke-data.csv file. The data contains 5110 rows and 12 features in total. The metadata of this datset is also available on Kaggle.

Dataset citation: FEDESORIANO. Stroke Prediction Dataset. Retrieved December 17, 2022 from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download.

2.2 Features

- id: unique identifier of a patient
- gender: "Male", "Female" or "Other"
- age: age of the patient
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- ever_married: "No" or "Yes"
- work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- Residence_type: "Rural" or "Urban"
- avg_glucose_level: average glucose level in blood
- smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- stroke: 1 if the patient had a stroke or 0 if not

2.3 Dateframe

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

Figure 2: Stroke Prediction Dataframe

2.4 Data Preprocessing

- Data Cleaning: I drop all the rows with missing value 'nan'. After doing so, the dataset contains 4909 rows. Except these missing values, the dataset is very clean and does not need further cleaning: The author has already performed necessary data cleaning.
- Data processing: I transform all the categorical data into dummy ones to enable machine learning models to process them.

2.5 Data FAIRness

- Findability: The stroke dataset is properly documented and has clear and accurate metadata. Hence, it is easy for others to discover and locate it.
- Accessibility: The stroke dataset is available in a format that is easy to use and that there are no barriers to accessing the data: It is totally free and liscensed by the author. Thus, the dataset can be easily accessible to those who need it.
- Interoperability: Using standardized formats and providing clear documentation about the data's structure and content, the dataset has a good interoperability, since it can be easily integrated with other datasets or tools.
- Reusability: The dataset provides clear documentation about the data's provenance, as well as any relevant ethical or legal considerations. Hence the dataset can be easily reused for multiple purposes.

2.6 Website Interface

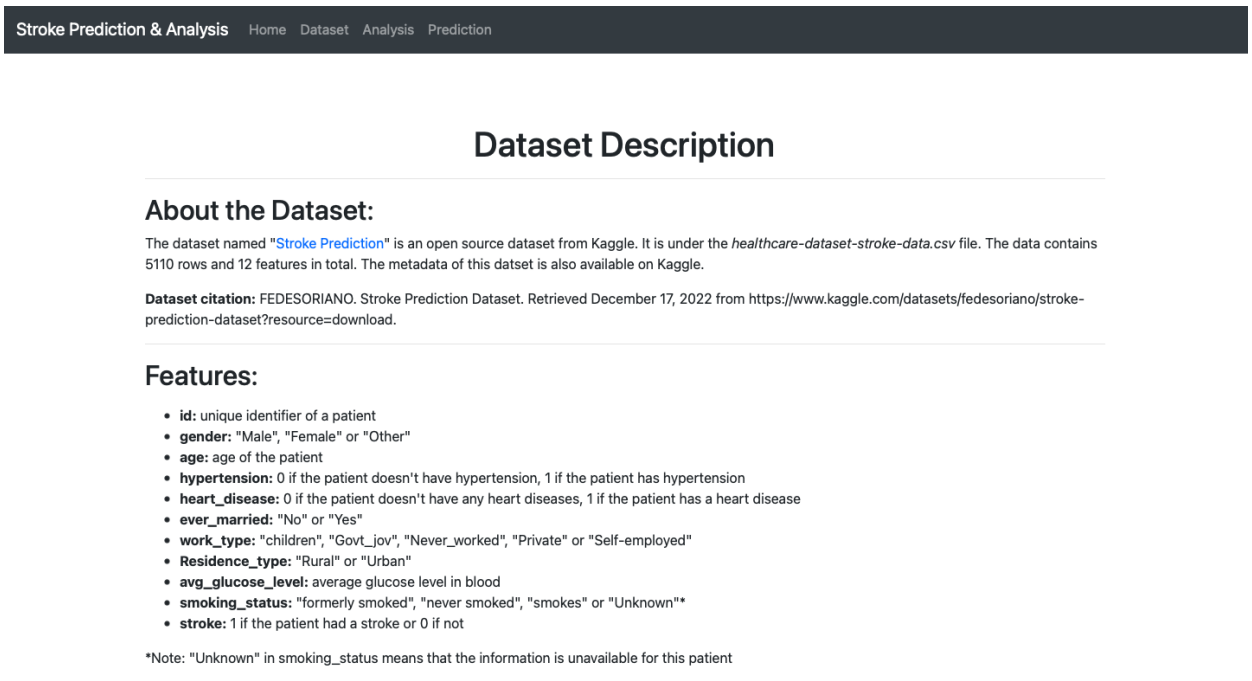


Figure 3: Webpage for Dataset Description

3 Exploratory Analysis

3.1 Summary Statistics

3.1.1 Statistical Analysis

	age	hypertension	heart_disease	avg_glucose_level	bmi
count	209.000000	209.000000	209.000000	209.000000	209.000000
mean	67.712919	0.287081	0.191388	134.571388	30.471292
std	12.402848	0.453486	0.394338	62.462047	6.329452
min	14.000000	0.000000	0.000000	56.110000	16.900000
25%	58.000000	0.000000	0.000000	80.430000	26.400000
50%	70.000000	0.000000	0.000000	106.580000	29.700000
75%	78.000000	1.000000	0.000000	196.920000	33.700000
max	82.000000	1.000000	1.000000	271.740000	56.600000

Figure 4: Summary statistics for patients with stroke

	age	hypertension	heart_disease	avg_glucose_level	bmi
count	4700.000000	4700.000000	4700.000000	4700.000000	4700.000000
mean	41.760451	0.083191	0.043191	104.003736	28.823064
std	22.268129	0.276201	0.203310	42.997798	7.908287
min	0.080000	0.000000	0.000000	55.120000	10.300000
25%	24.000000	0.000000	0.000000	76.887500	23.400000
50%	43.000000	0.000000	0.000000	91.210000	28.000000
75%	59.000000	0.000000	0.000000	112.432500	33.100000
max	82.000000	1.000000	1.000000	267.760000	97.600000

Figure 5: Summary statistics for patients without stroke

From the statistics, it is clear that:

- The mean age for people with stroke is much higher than those without stroke. Patients with stroke tends to be 26 years elder than the healthy on average.
- Patients with stroke have higher probability to have hypertension than healthy people.
- Patients with stroke have slightly higher probability to have heart disease than healthy people.
- Patients with stroke tends to have a higher average glucose level than healthy people.

In conclusion, elder people with other chronic diseases have a higher possiblility to have stroke.

3.1.2 Data Distribution and Outliers Analysis

In our dataset, there are three numberic features. The above figure are histogram and barplot, which shows the distribution of these data. From the figure:

- There is no outlier for age.
- There are a lot of outliers for avg glucose level and bmi. All the outlier are high values.

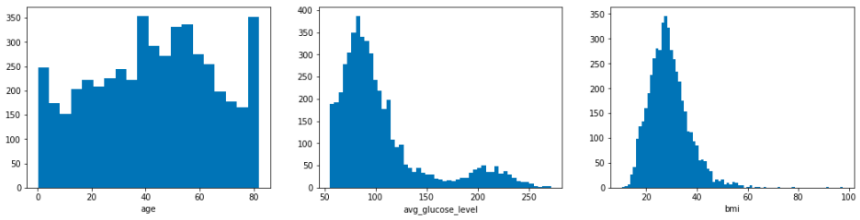


Figure 6: Histogram for features

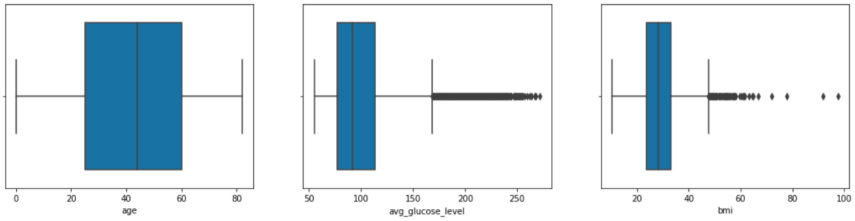
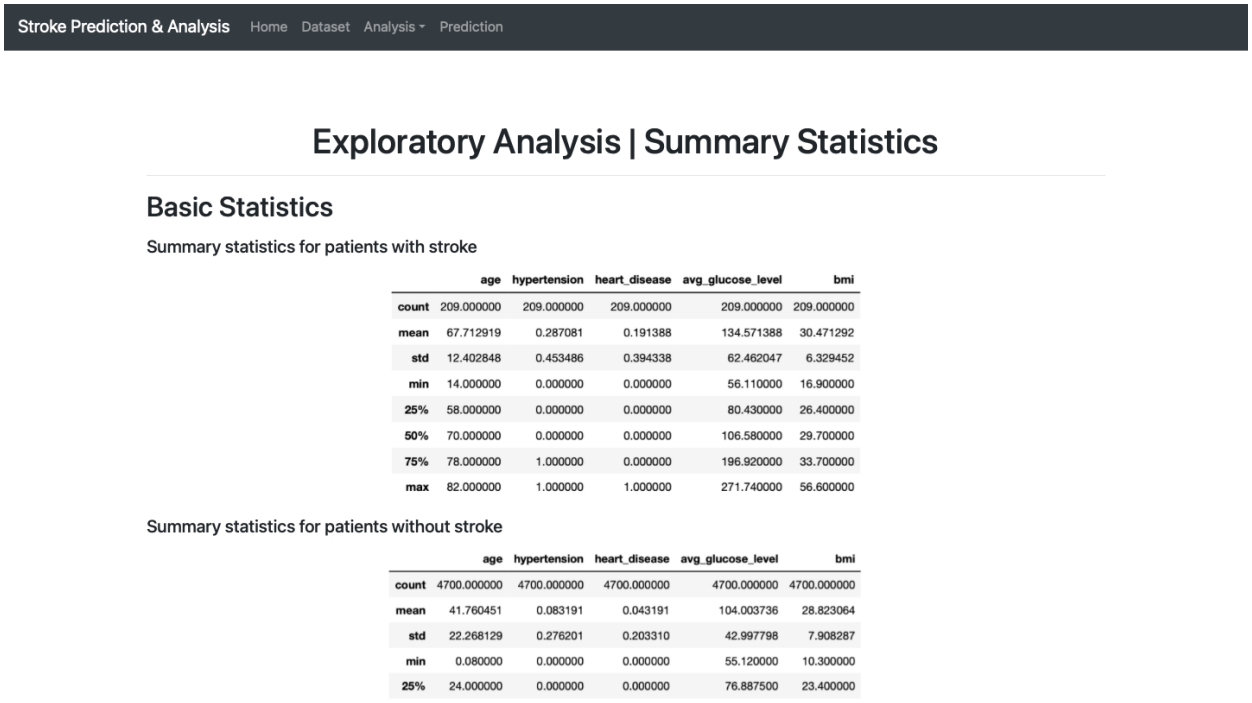


Figure 7: Bar plot for features

3.1.3 Website Interface



	age	hypertension	heart_disease	avg_glucose_level	bmi
count	209.000000	209.000000	209.000000	209.000000	209.000000
mean	67.712919	0.287081	0.191388	134.571388	30.471292
std	12.402848	0.453486	0.394338	62.462047	6.329452
min	14.000000	0.000000	0.000000	56.110000	16.900000
25%	58.000000	0.000000	0.000000	80.430000	26.400000
50%	70.000000	0.000000	0.000000	106.580000	29.700000
75%	78.000000	1.000000	0.000000	196.920000	33.700000
max	82.000000	1.000000	1.000000	271.740000	56.600000

	age	hypertension	heart_disease	avg_glucose_level	bmi
count	4700.000000	4700.000000	4700.000000	4700.000000	4700.000000
mean	41.760451	0.083191	0.043191	104.003736	28.823064
std	22.268129	0.276201	0.203310	42.997798	7.908287
min	0.080000	0.000000	0.000000	55.120000	10.300000
25%	24.000000	0.000000	0.000000	76.887500	23.400000

Figure 8: Webpage for Summary Statistics

3.2 Univariate Analysis

3.2.1 Analysis Questions

- Which age/gender has the highest probability to have stroke?
- How avg glucose level and bmi related to stroke?

3.2.2 Age Distribution

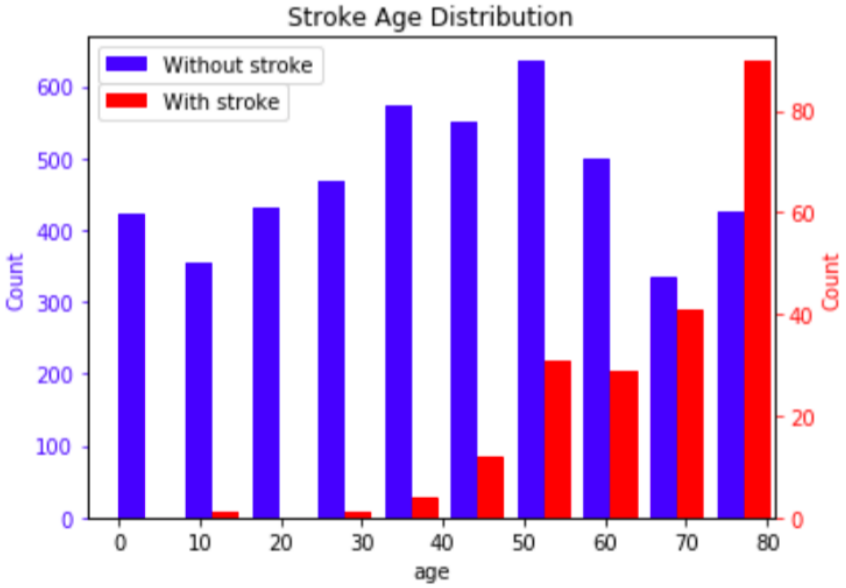


Figure 9: Histogram for age distribution

The larger the age is, the more possible a person have stroke.

3.2.3 Gender Distribution

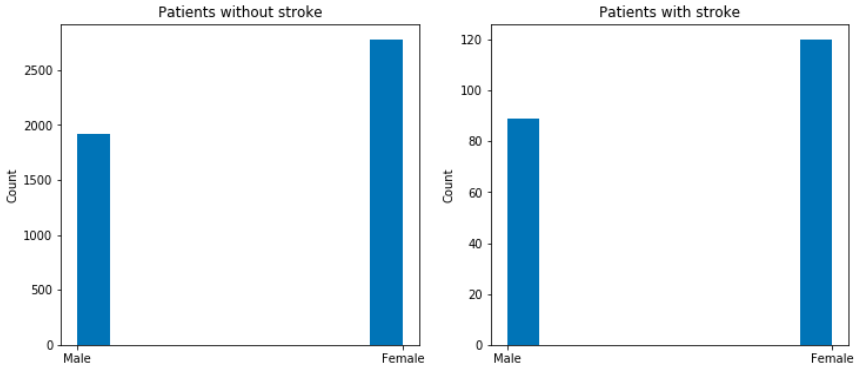


Figure 10: Histogram for gender distribution

The dataset contains more female patients than male ones. By comparing the proportion of gender within different groups, it can be concluded that there is no strong relationship between gender and stroke.

3.2.4 Glucose Distribution

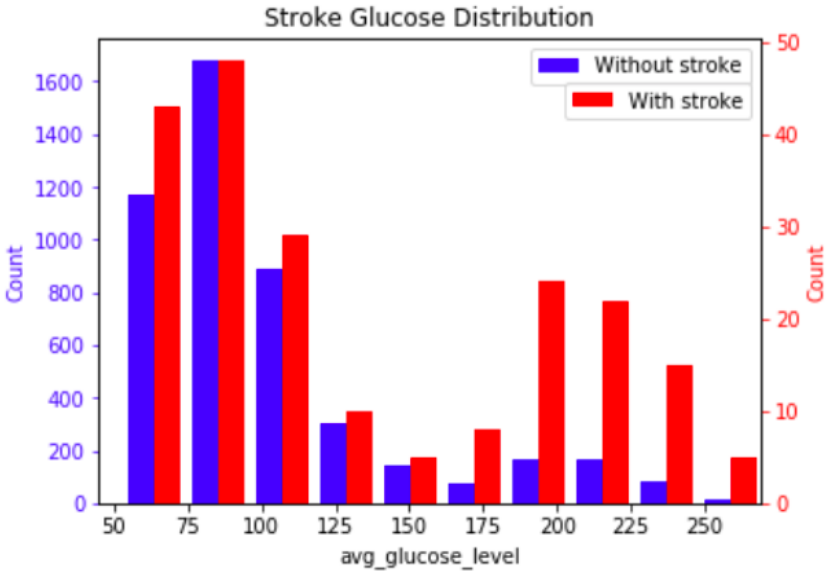


Figure 11: Histogram for glucose distribution

From the histogram, a higher glucose level do suggest a higher probability to have stroke. However, for patients with regular average glucose levels, the probability of having stroke won't decrease.

3.2.5 BMI Distribution

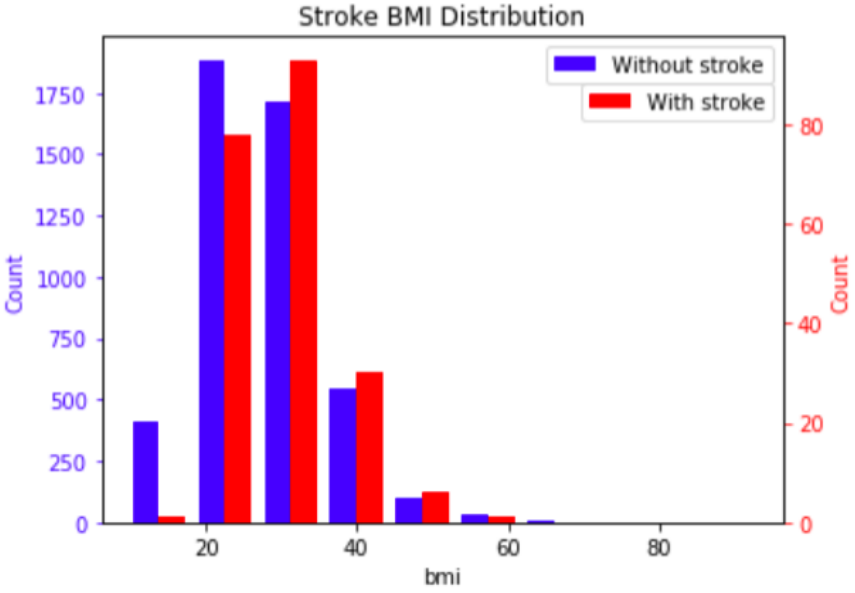


Figure 12: Histogram for BMI distribution

The histogram suggests that stroke patients tend to have a higher bmi. There exists a weak correlation between bmi and stroke.

3.2.6 Website Interface

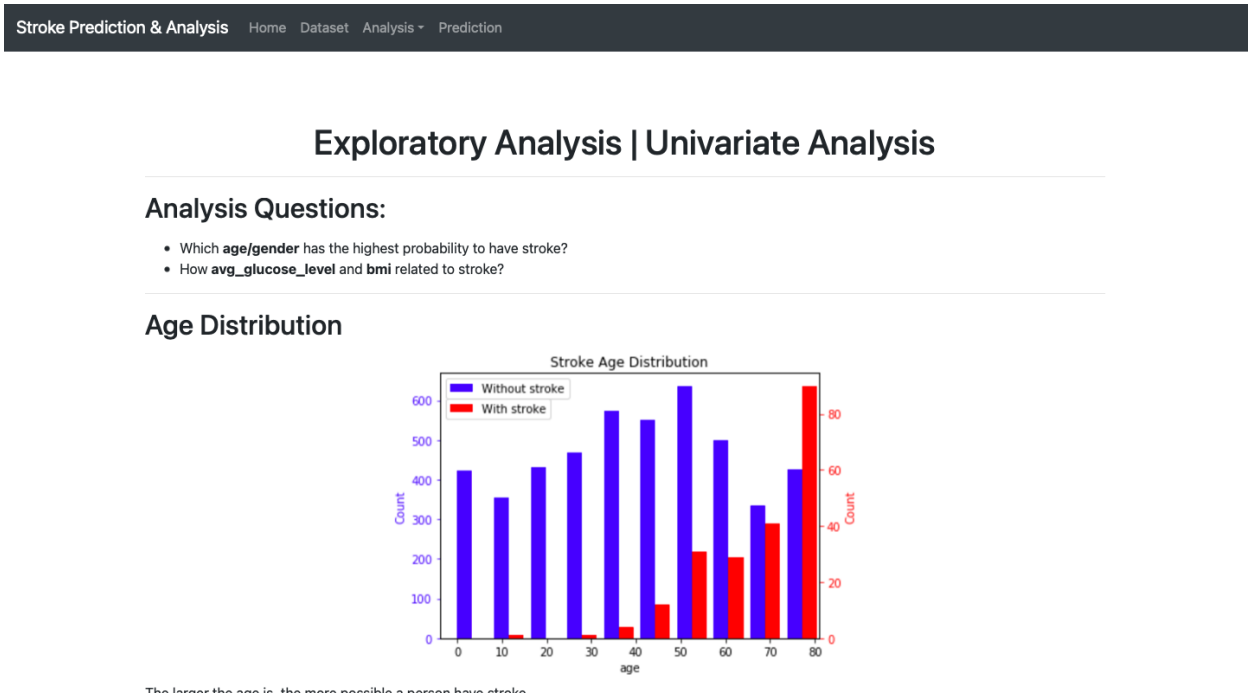


Figure 13: Webpage for Summary Statistics

3.3 Bivariate Analysis

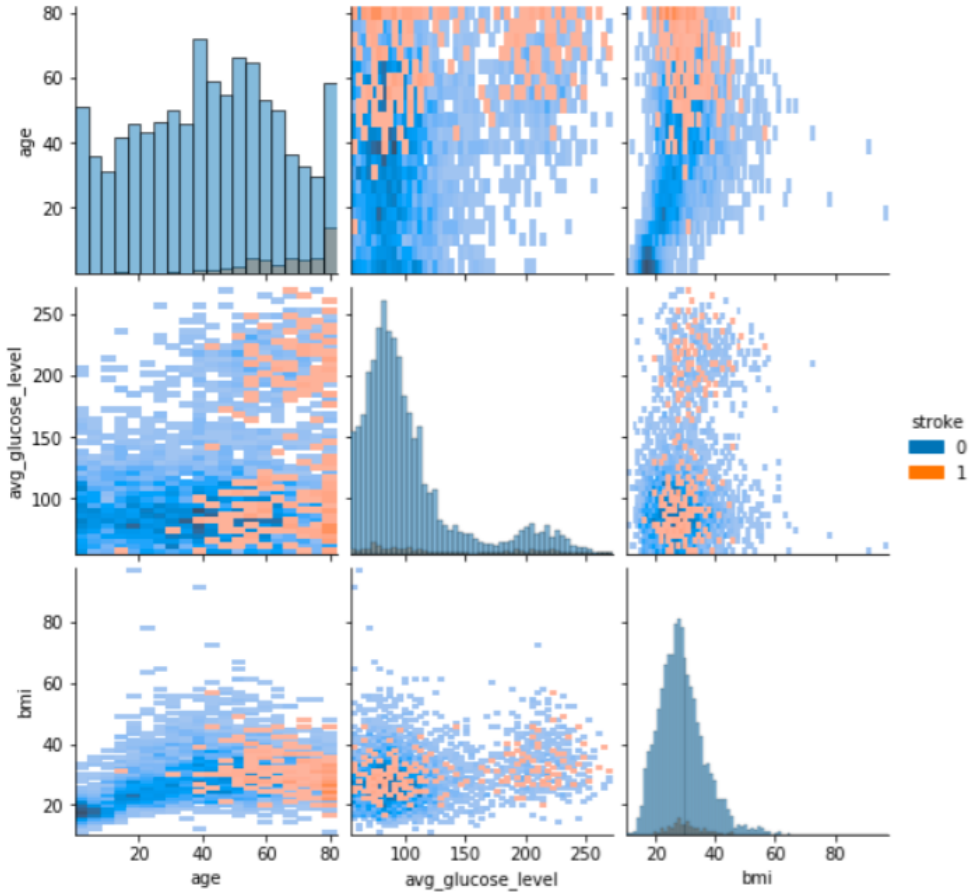


Figure 14: Summary plot for features

The above figure is the pairplot of three numeric features: age, avg glucose level and bmi. Patients with higher age are more likely to have stroke. A higher average glucose level and a larger bmi are more likely to result in stroke.

3.3.1 Website Interface

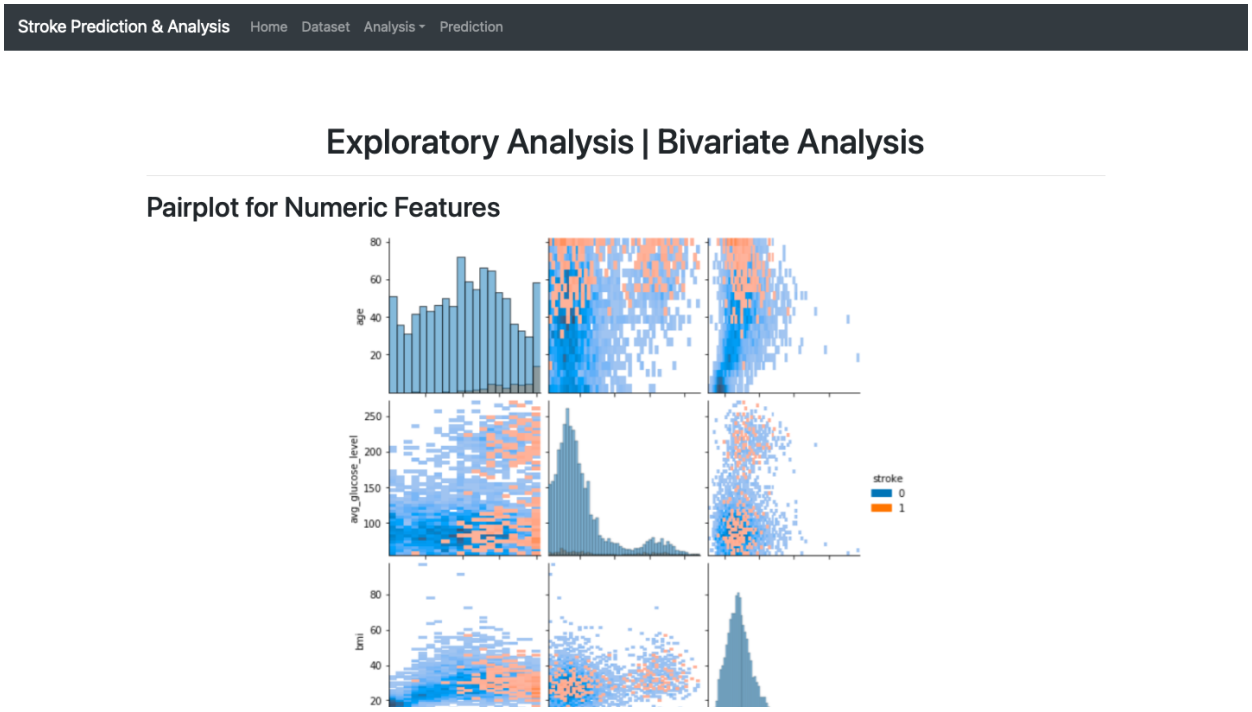


Figure 15: Webpage for Bivariate Analysis

3.4 Feature Correlation

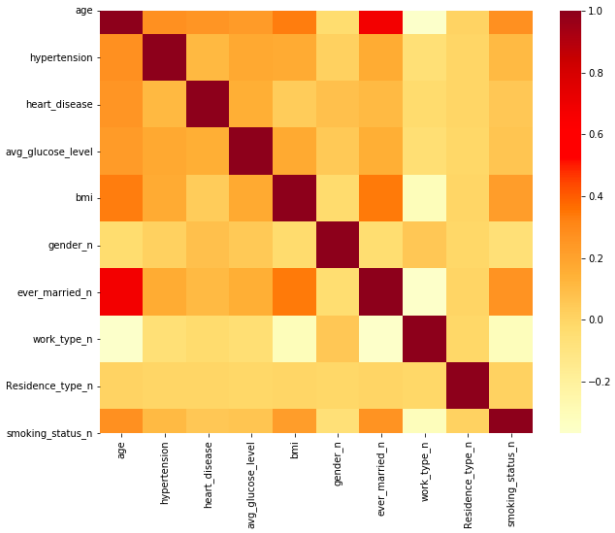


Figure 16: Correlation matrix for all features

First, I transform all the categorical features into numeric ones.

Then I plot the correlation matrix of the 10 features. The heatmap shows the Pearson correlation coefficients between the features in my dataset. The relationships between the features can then be identified and how they may affect the target variable be understood.

The Pearson correlation coefficient is a measure of the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a strong negative relationship, 0 indicates no relationship, and 1 indicates a strong positive relationship. A correlation matrix can help identify which features are highly correlated with each other and which are not.

If two features are highly correlated, it may be beneficial to remove one of them from the model to avoid overfitting and improve the model's performance. From the result, it is shown that there does not exist two features that are highly correlated. Thus I keep all the 10 features to train the models.

3.4.1 Any surprise

The feature correlation are small, so I keep all these features.

3.4.2 Website Interface

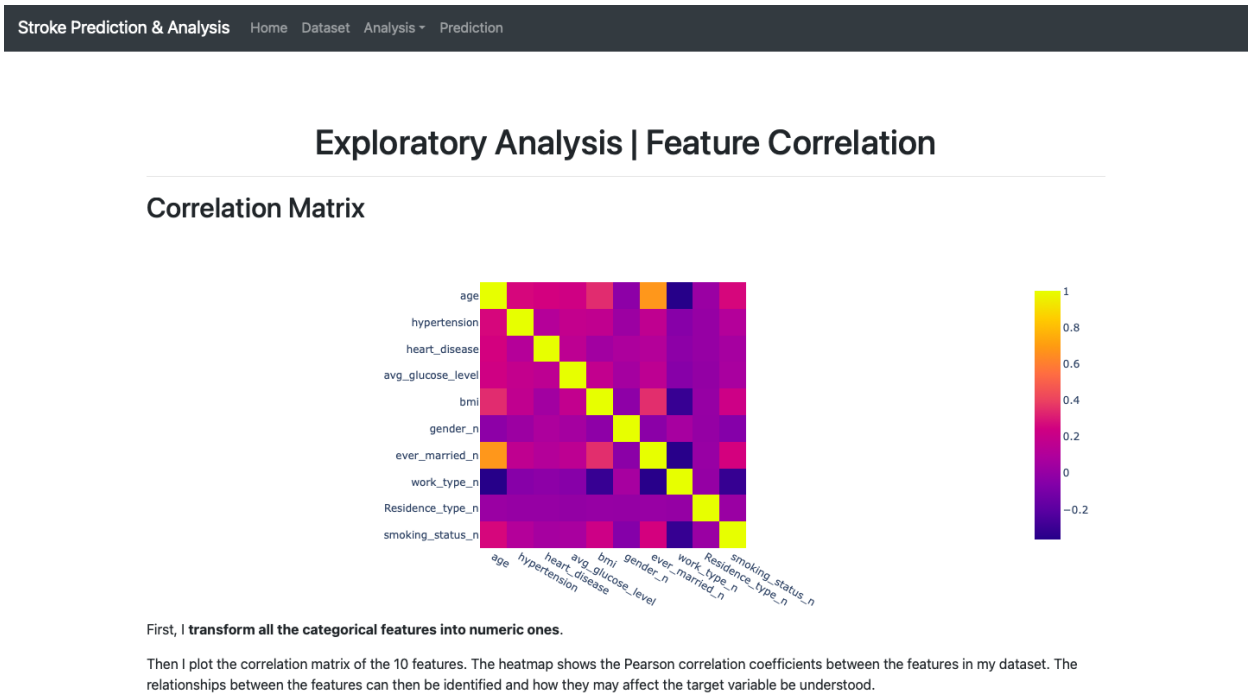


Figure 17: Webpage for Feature Correlation

4 Prediction

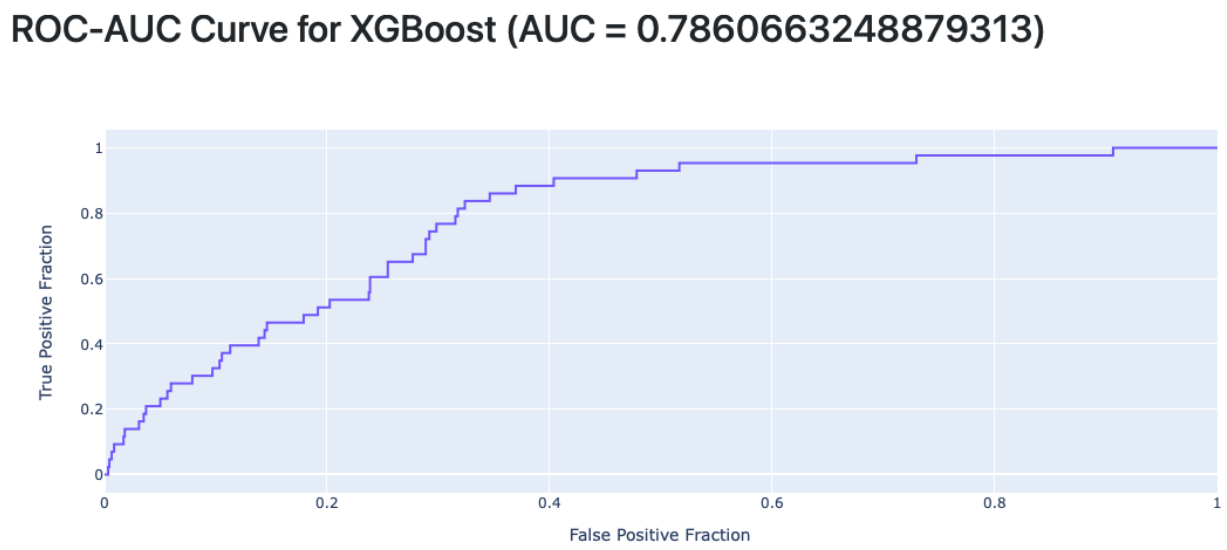
I choose two model for prediction. One is XGBoost.

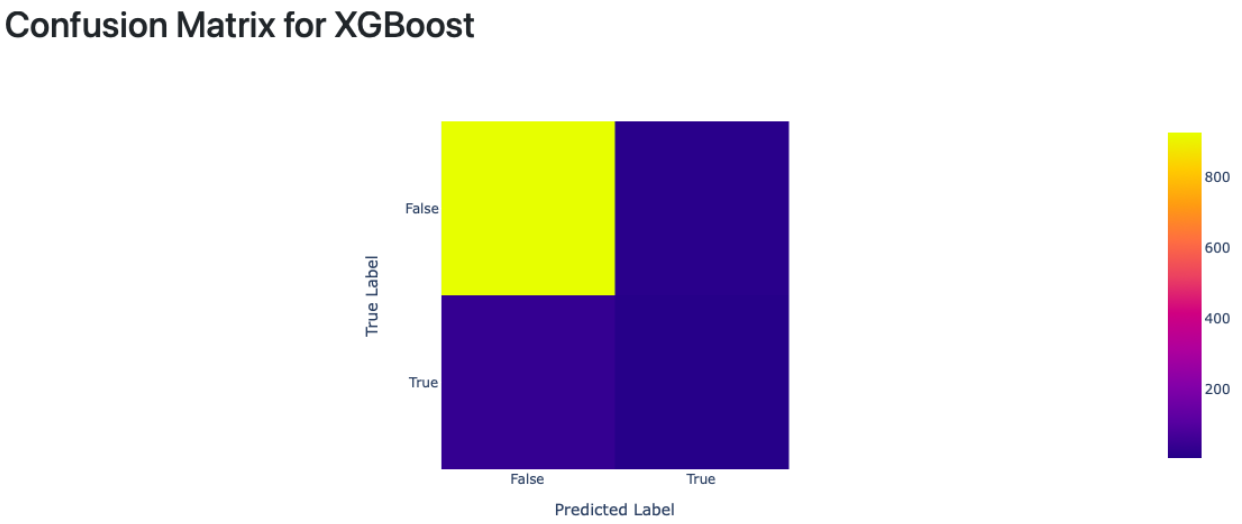
4.1 Performance of XGBoost

XGBoost (eXtreme Gradient Boosting) is a popular and efficient open-source implementation of the gradient boosting algorithm for machine learning.

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The main idea behind gradient boosting is to train weak models sequentially, each trying to correct the mistakes of the previous model.

Overall, XGBoost is a powerful and flexible tool for implementing gradient boosting and is well-suited for a wide range of machine learning tasks.





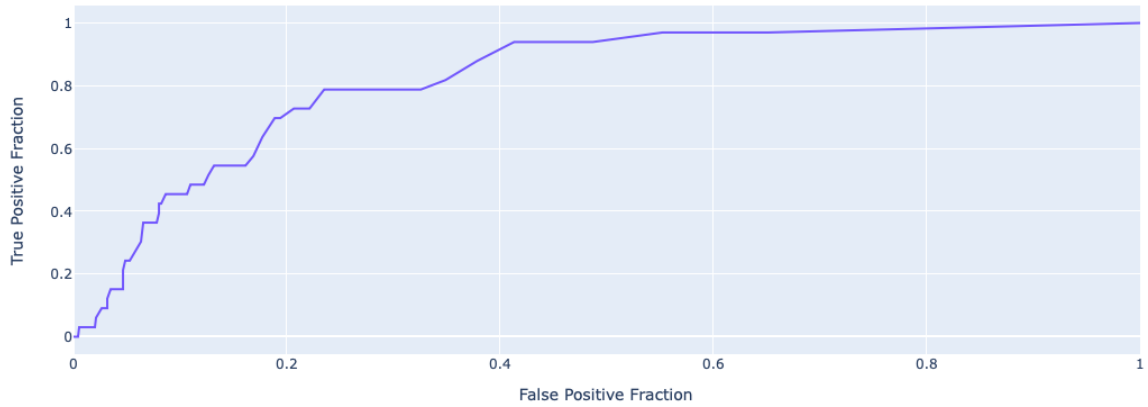
4.2 Performance of Random Forest

Random Forest is a popular and powerful ensemble machine learning algorithm that is used for classification and regression tasks.

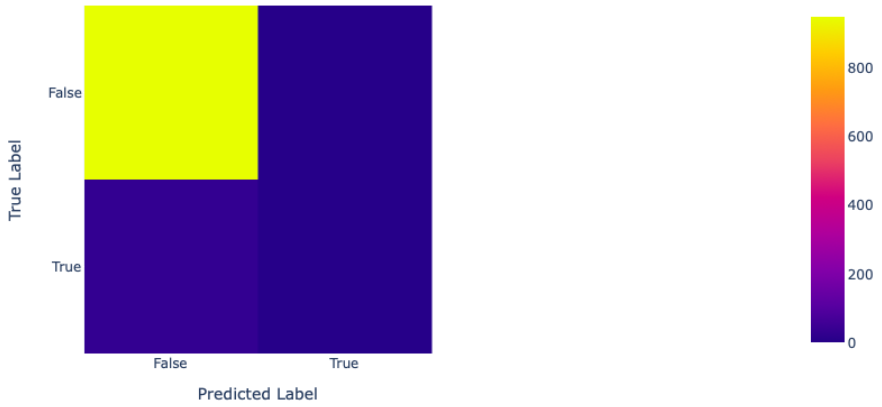
Random Forest is a flexible and easy-to-use algorithm that can handle a large number of input features and can deal with missing values and categorical variables automatically. It is also relatively resistant to overfitting, due to the way it combines multiple decision trees.

Overall, random forest is a widely used and robust machine learning algorithm that is well-suited for many applications.

ROC-AUC Curve for Random Forest (AUC = 0.8231312066928504)



Confusion Matrix for Random Forest



4.2.1 Website Interface

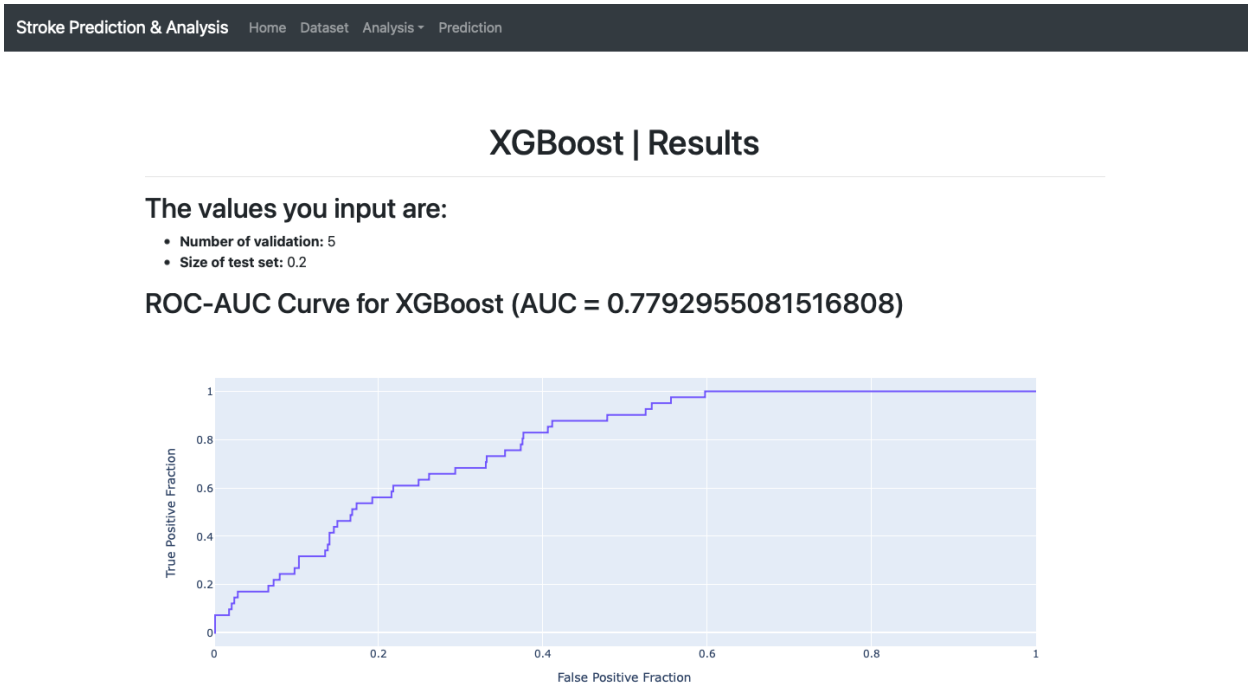


Figure 18: Webpage for Feature Correlation

5 Conclusion

In conclusion, in this final project, I analyze stroke prediction dataset and use XGBoost and Random Forest to make prediction. The results are satisfactory. I also made a website which have all the information on it.