Project Title: Machine Learning for Brain Stroke

Problem Statement

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Stroke is a condition that occurs when the blood supply to the brain decreases due to a blockage (ischemic stroke) or rupture of a blood vessel (hemorrhagic stroke). Without blood, the brain will not get oxygen and nutrients, so the cells in the affected brain area will soon die. Prompt treatment can minimize the level of damage to the brain and the possibility of complications. Over the past few decades, Machine Learning (ML) delivers an accurate and quick predict outcome and it has become a powerful tool in health settings, offering personalized clinical care for stroke patients. Machine learning methods are flexible prediction algorithms with potential advantages over conventional regression and scoring system. XGBoost has been widely recognized in the number of machine learning and data mining challenges. In this study, we will explore Stroke Prediction Dataset using XGBoost and compare with various other machine learning models and techniques to predict status of stroke.

Data Source

The data we are using is from Stroke Prediction Dataset. The dataset is from Kaggle [1]. Each row in the data provides relevant information about the patient. The data types are mixed and include strings, integer and floats. Each row represents an individual record. The dependent variables are age, hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, smoking status and stoke. The dataset contains data associated with 5110 patients. The mean age of adult patients is 43.22 years, 41.4% patients are male, the hypertension ratio is 9.7%, The heart disease ratio is 5.4%, The stroke ratio in this dataset is 4.9%. The brief data summary is in the Table 1. Figure 1 show the stroke distribution by age. It clearly shown that elder people can get the stroke more often than the young people. The other data summary graphs and tables are in Appendix. We observed that people that are ever smoke (both formerly and presently) have a relatively high chance to have a stroke in Appendix Figure 4. The Appendix Figure 2 show that people that have ever had both heart disease and hypertension are most likely to have a stroke. On the other hand, people that never have those diseases tend to not have a stroke too.

Table 1: Summary statistics of the study population

Features	Statistics (mean ± std)
Age	43.22 ± 22.61
Gender	Male: 41.4% Female: 58.6%
Hypertension ratio	9.7%
Heart disease ratio	5.4%
Ever married ratio	65.6%
Avg Glucose Level	106.14 ± 45.28
BMI	28.89 ± 7.85
Stroke ratio	4.9%

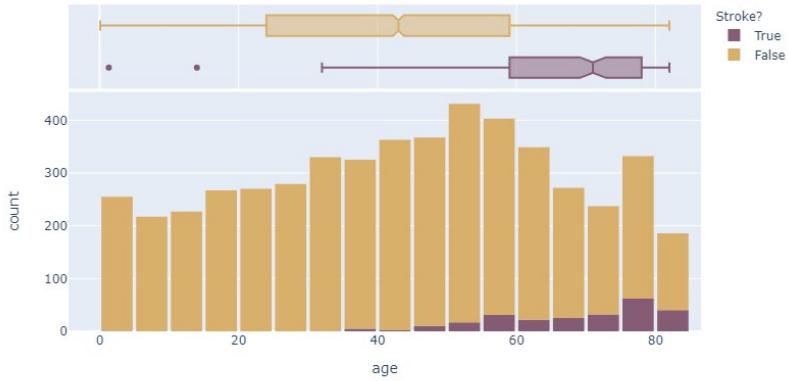

Figure 1. The stroke distribution by age.

Methodology

EXtreme Gradient Boosting (XGBoost) is a machine learning technique with the remarkable features of processing the missing data efficiently and flexibly and assembling weak prediction models to build an accurate one [2]. As an open source package, XGBoost has been widely recognized in a number of machine learning and data mining challenges, for example, 17 solutions used XGBoost among the 29 challenge winning solutions published at Kaggle's blog in 2015 and the top-10 winning teams used XGBoost in KDD Cup 2015. novel machine learning techniques have demonstrated improved predictive performance compared to traditional prediction methods. In this project. We attempted to compare the performance of machine learning (XGboost) model with several others traditional prediction models (Random forest, Logist regression, Dicision Tree, Gradient Boosting AdaBoost) [3,4,5]. Methodology flowchart of this project is shown in Figure 2, and the set-ups of all these methods will be introduced in details in their respective parts.

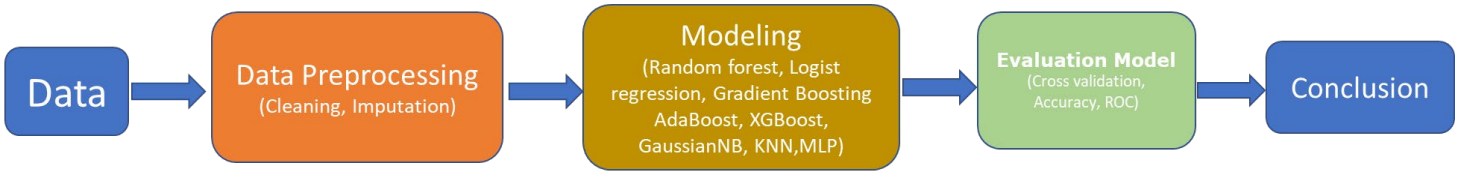

Figure 2: Methodology Flowchart

Data Preprocessing

Missing value and Data Imputation

Based on the Figure 2. We have found that there are 201 missing values in the bmi and 1544 unknown values for smoking status. Other variables do not have missing values. Total missing rate is 2.8%. Based on the Figure 3. We can see that the missing variables did not show clear association with other variables and therefore we can assume this missingness is MCAR (missing completely at random). The final missing dataset were imputed for the bmi values with the median and the smoking status with the most frequent value. There are no features that are highly correlated with other features based on the correlation heat map as shown in Figure 1.

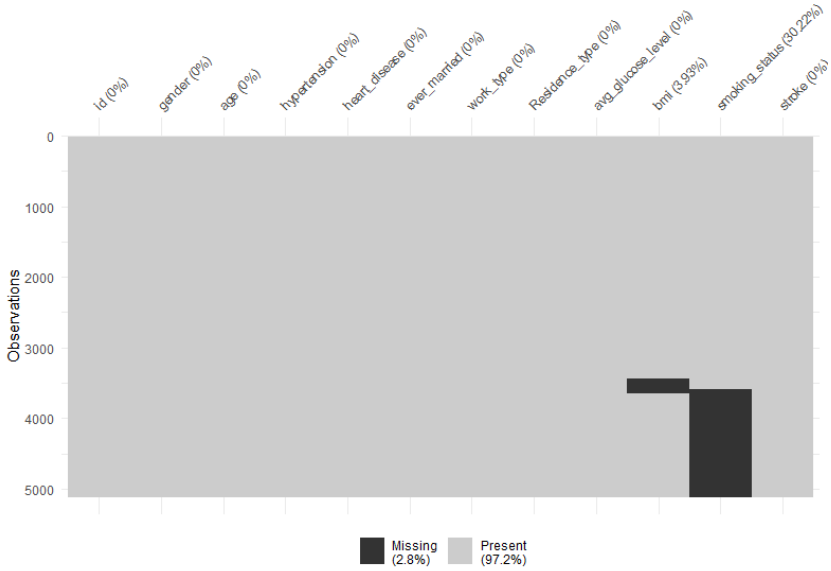

Figure 3. The missing value distribution by features.

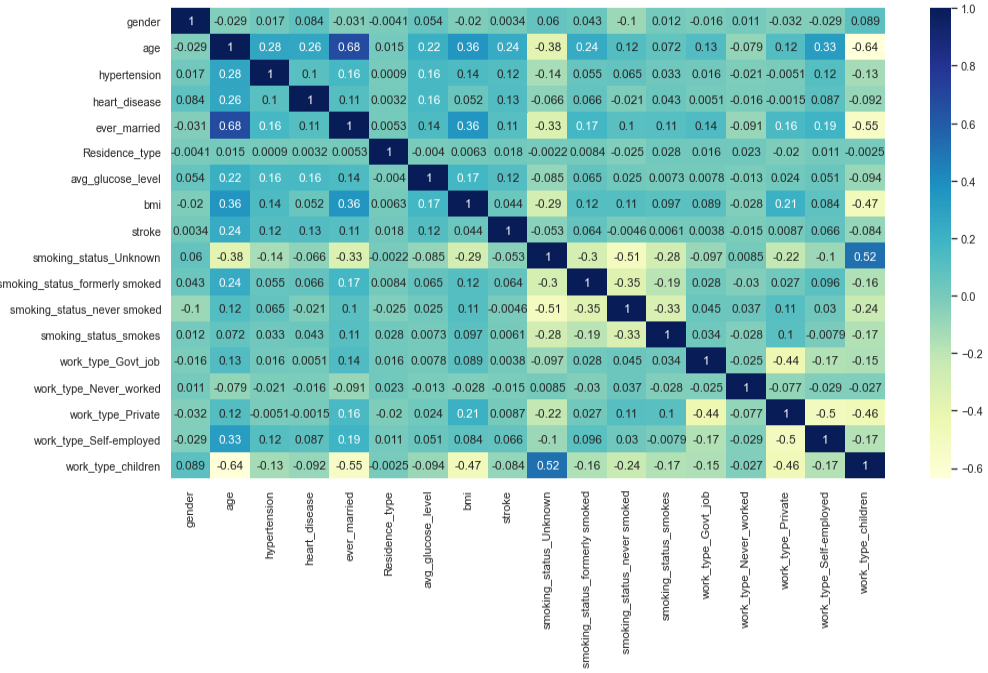

Figure 4. The correlation between the features.

Evaluation

The dataset will be split into 80% training set and 20% test set. Hyperparameter tuning was done on 5-fold CV of the training set and the final evaluation of model performance was done on the test set. According to the highly imbalance of this dataset in Figure 5, A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. Synthetic Minority Oversampling Technique (SMOTE) is a very popular oversampling method that was proposed to improve

random oversampling with low-dimensional data. So, we will apply SMOTE (Synthetic Minority Oversampling Technique) to oversample the dataset and then the models can learn more efficiently [6].

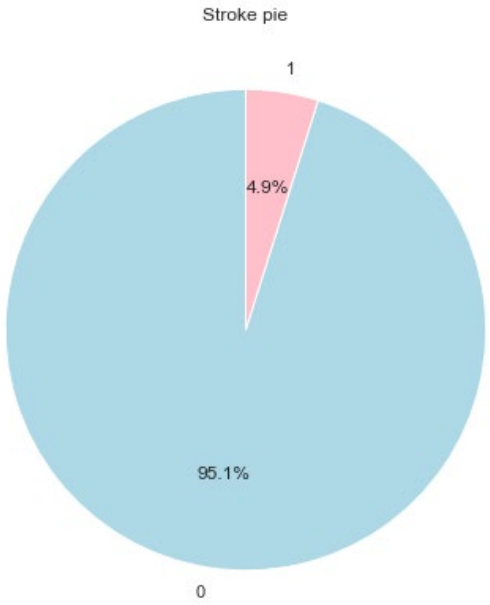

Figure 5. The stroke Pie distribution.

Multiple machine learning models (Random Forest, Logistic Regression, GradientBoost, AdaBoost, XGBoost, GaussianNB, KNN, MLP) in the study were evaluated and compared using the Area under the Receiver Operating Characteristic curve (ROC) on the test set. The ROC curve is the true positive rate against the false positive rate at various threshold settings. ROC provides a single measure of the diagnostic ability of a binary classifier. Apart from the primary metric of AUROC, accuracy, RMSE, precision, recall and F1-score were also used during the model testing stage to provide a full picture of model performance.

Results

The results are based on the 5-fold cross validation. Based on the ROC curve graph in Figure 5, we found that KNN is worst method to predict the stroke in this dataset. Logistic regression, GradientBoost, AdaBoost Random Forest and XGBoost are performing best compared with the rest of methods. Based on the Accuracy and RMSE in the Table 2. XGBoost method has best accuracy 0.9325 and lowest RMSE value 0.2598. The random forest had a resulting RMSE of 0.2635 and an accuracy score of 0.9305. While Logistic regression models have coefficients that we can directly interpret and, in this case, are faster to train and test, we opted to select the random forest for its increase in accuracy. The precision, recall and F1-score are shown in the table 3.

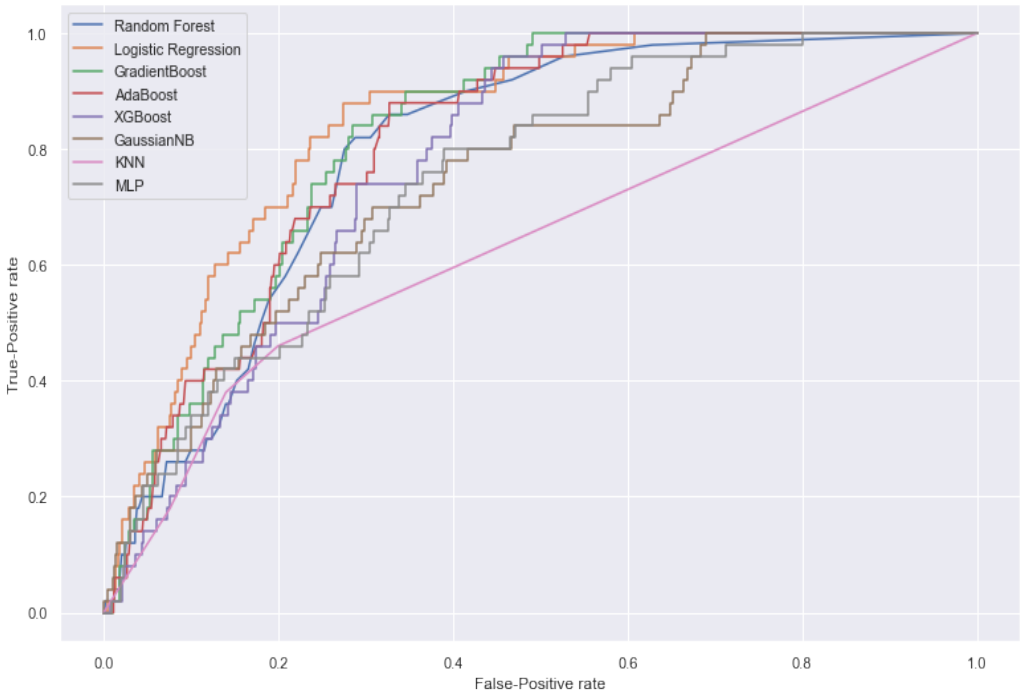

Figure 5. ROC trend by eight machine learning methods

Table 2: Accuracy and RMSE for the models

Methods	Accuracy score	RMSE
Random Forest	0.9305	0.2636
Logistic Regression	0.7329	0.5168
GradientBoost	0.8992	0.3175
AdaBoost	0.8366	0.4042
XGBoost	0.9325	0.2598
GaussianNB	0.3023	0.8353
KNN	0.8464	0.3919
MLP	0.864	0.3688

Table 3: The precision, recall and F1-score for the model

Model		precision	recall	f1-score
Random Forest	0	0.95	0.97	0.96
	1	0.14	0.08	0.1
Logistic Regression	0	0.99	0.73	0.84
	1	0.13	0.82	0.23
GradientBoost	0	0.97	0.92	0.95
	1	0.22	0.4	0.28
AdaBoost	0	0.97	0.85	0.91
	1	0.16	0.56	0.25
XGBoost	0	0.95	0.98	0.96
	1	0.09	0.04	0.05

GaussianNB	0	1	0.27	0.42
	1	0.07	1	0.12
KNN	0	0.96	0.87	0.92
	1	0.13	0.38	0.19
MLP	0	0.96	0.89	0.93
	1	0.14	0.36	0.21

Discussion

This study demonstrated that the use of machine learning models can accurately predict the stroke patients. various machine learning algorithms have been investigated for prediction performance. The AUCs, accuracy score and MSRE we developed have demonstrated the benefit of using a XGboost model as opposed to the other machine learning for prediction of stroke. Our results have showed that XGBoost has outperform the other methods based on all these three criterions. But there are several limits in XGBoost methods. For instance, the features selected were according to clinical experience but not algorithm; the representativeness of features may not clear in stroke and some important dynamic features were not included; besides, there were no validations for the XGboost model and no traditional regression analysis was used as a control.

The strength of this study was mainly that it was to predict the stroke using the XGBoost model. And compared to traditional regression analysis. We must acknowledge some other limitations of our study and it may provide a base for potential improvement: firstly, because the data come from only one database and incidence of stroke is very small; secondly, further exploration for the database was not performed, which may lead to the abandonment of some key variables; Thirdly, the proposed model was not designed to be validated by developing set from the database or our clinical data. Even so, we believe that the proposed model may contribute to further our understanding of the stoke.

Conclusion

In this project, we investigate an interesting topic of predicting the Stroke by implementing 8 machine learning models (Random Forest, Logistic Regression, GradientBoost, AdaBoost, GaussianNB, KNN, MLP, XGBoost)., lasso regression, SVM, decision tree regressor and random forest regressor). Based on the cross validation results. XGboost generally performs the best on predicting all awards. Using machine learning technique by XGboost, more significant prediction model can be built. This XGboost model may prove clinically useful and assist clinicians in tailoring precise management and therapy for the patients with Stroke.

Reference

1. Stroke Prediction Dataset, https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
2. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining-KDD 2016, San Francisco, CA, USA; 2016. p.785–94.
3. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
4. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2000;29(5):1189–232.
5. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
6. Blagus, R., Lusa, L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106 (2013). https://doi.org/10.1186/1471-2105-14-106

Appendix

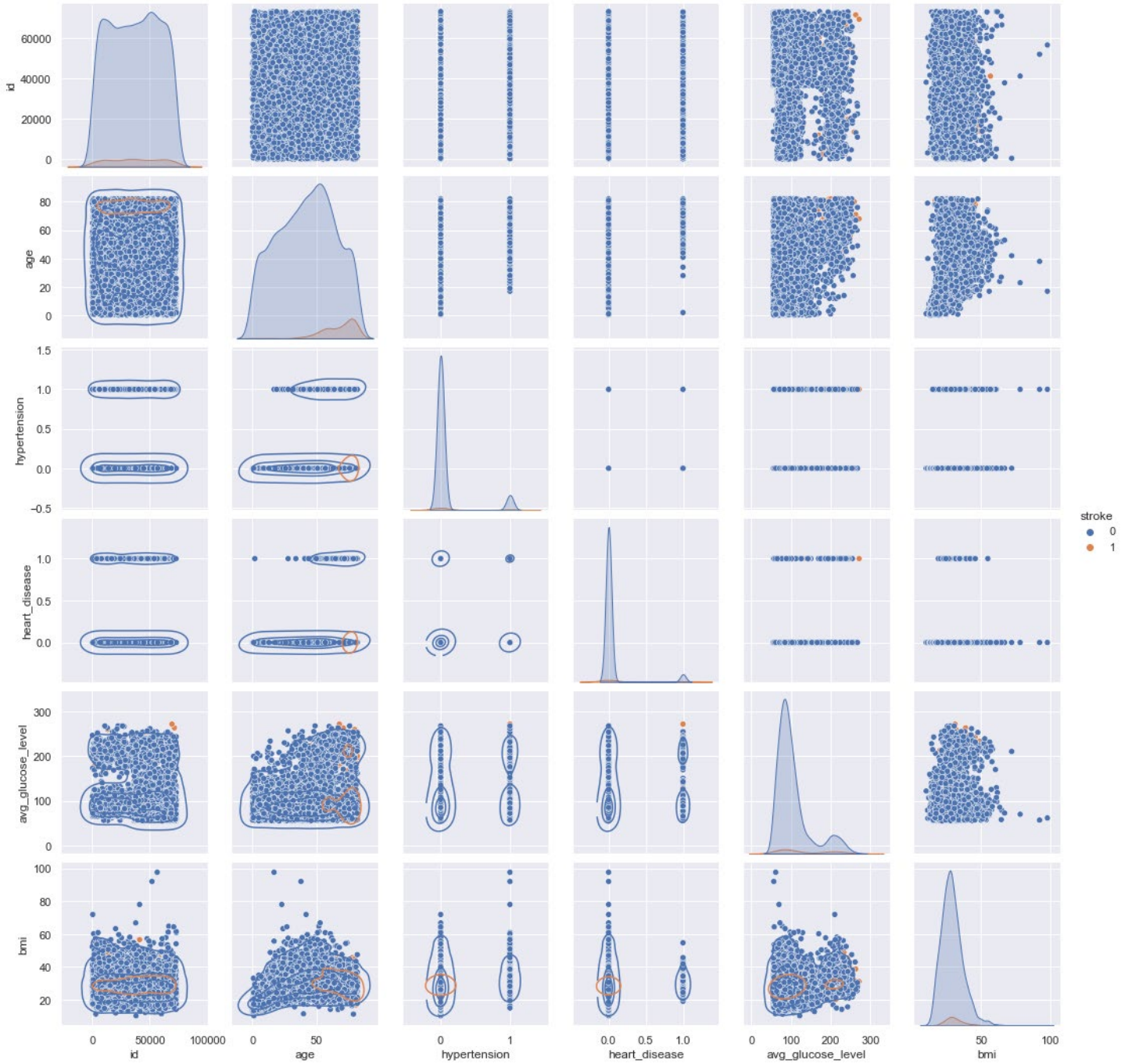

Figure 1. The KDE density graph between the features (id, age, hypertension, heart_disease, avg_glucose_level, bmi).

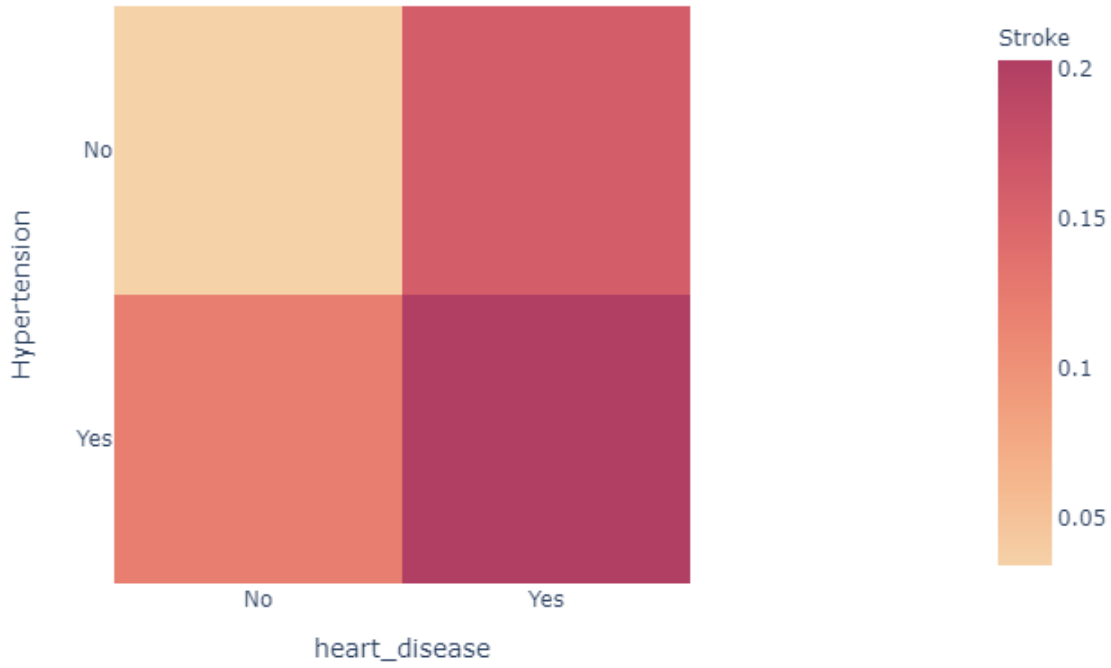

Figure 2. The Stoke probabilities by heart disease and Hypertension

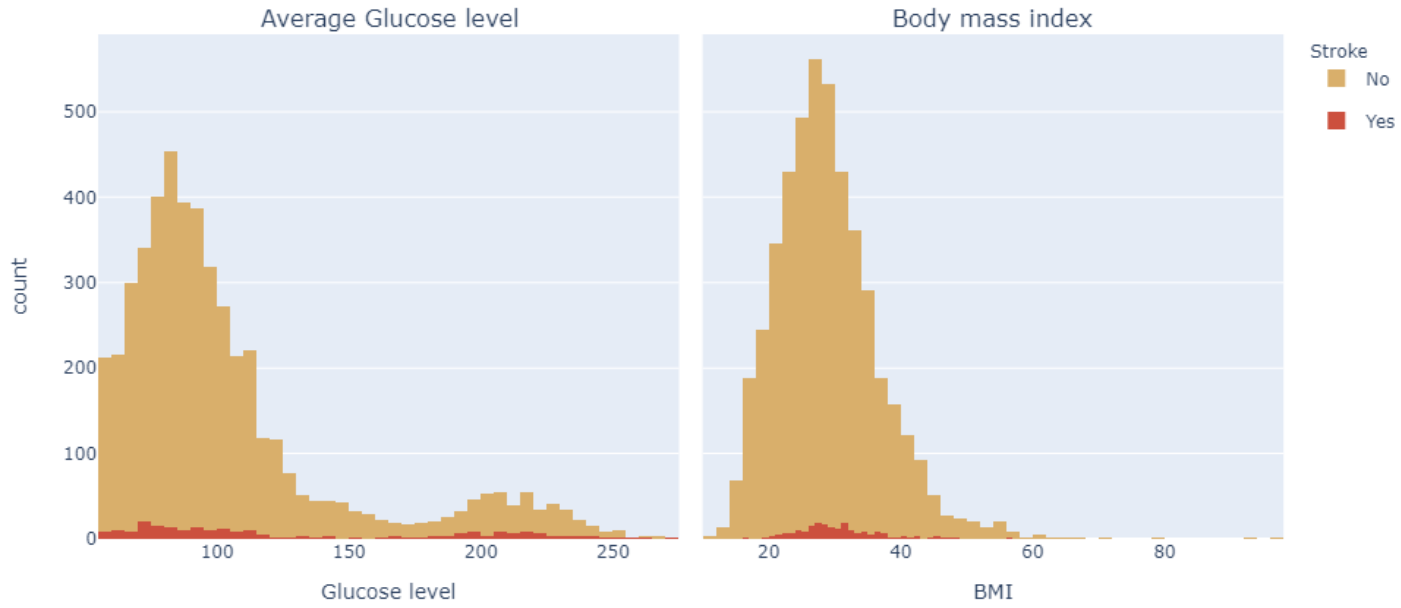

Figure 3. The Stoke distribution by average glucose level and BMI.

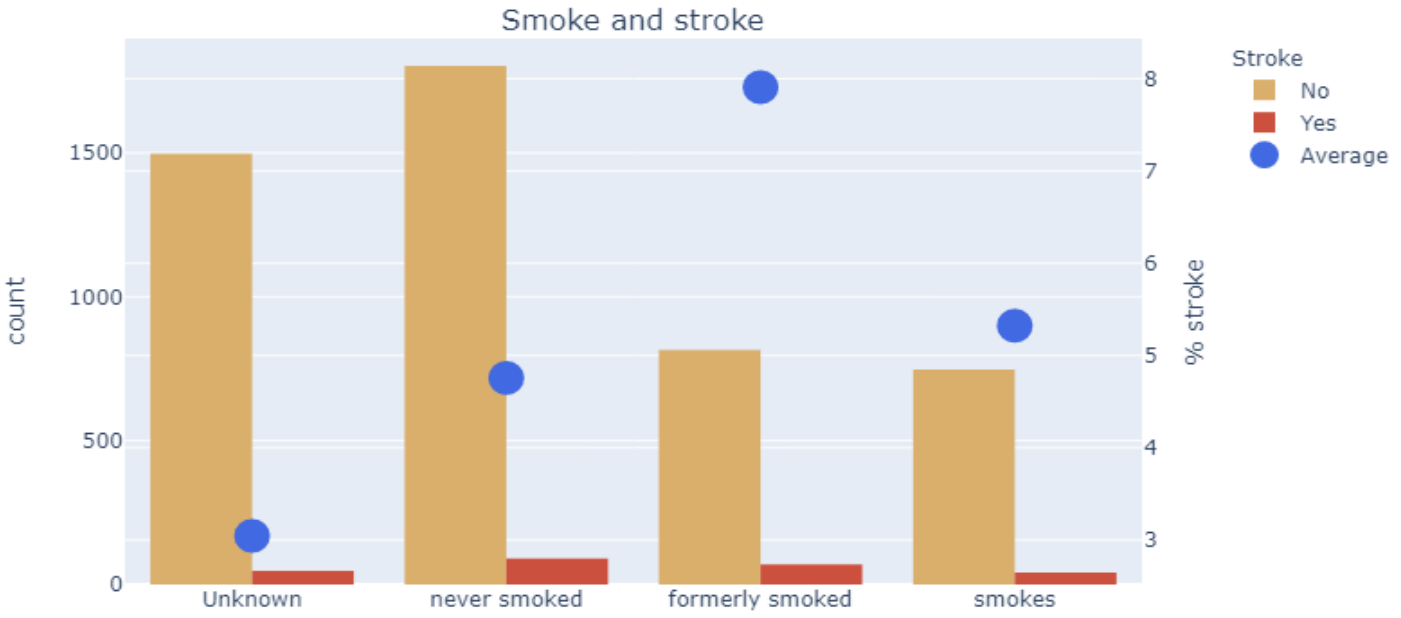

Figure 4. The Stoke distribution by smoke status.

Features	Category	sum	count	Stroke mean
ever_married	No	29	1757	0.016505
	Yes	220	3353	0.065613
work_type	Govt_job	33	657	0.050228
	Never_worked	0	22	0
	Private	149	2925	0.05094
	Self-employed	65	819	0.079365
	children	2	687	0.002911
Residence_type	Rural	114	2514	0.045346
	Urban	135	2596	0.052003

Table 1: The stroke distribution by features (married, work type and Residence type).