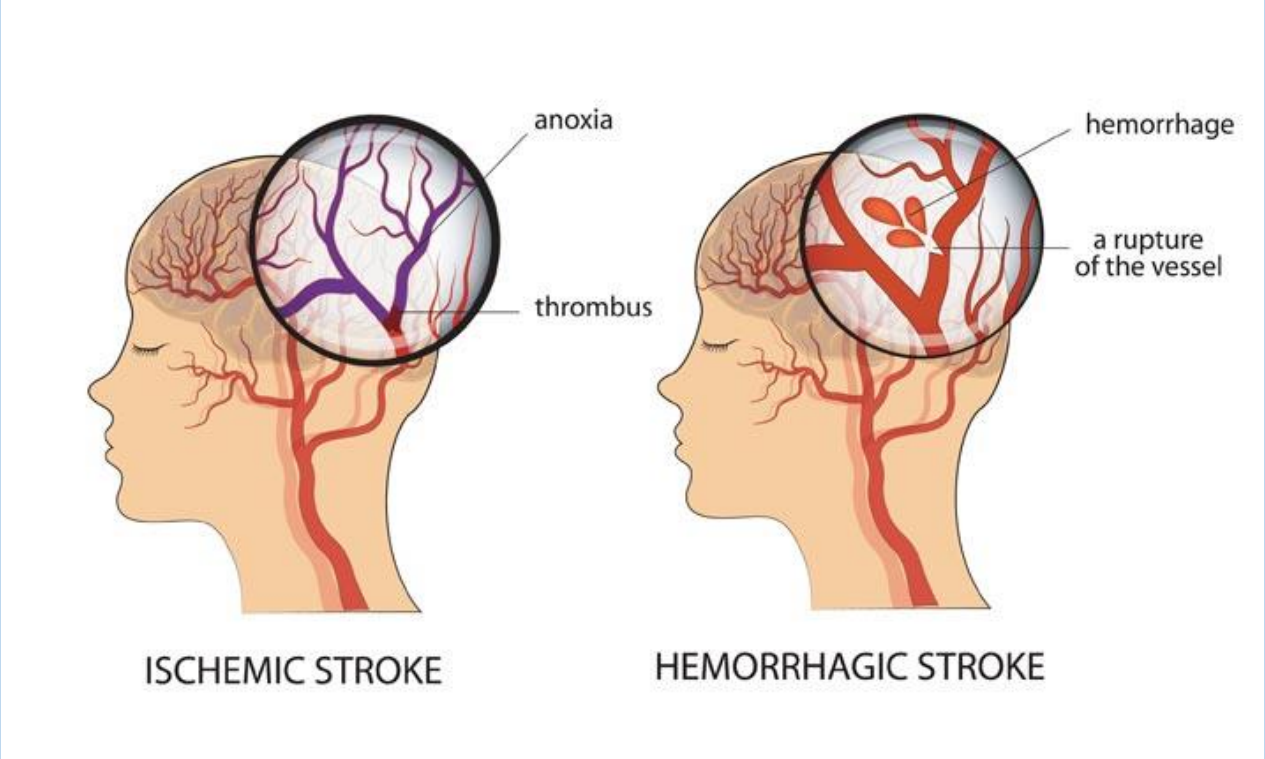
Stroke Prediction

Yuliya Selevich

Data Science Capstone Project

The problem

- Stroke is the No. 5 cause of death and a leading cause of disability in the United States
- Stroke-related costs in the United States came to nearly **$53 billion** between 2017 and 2018. This total includes the cost of health care services, medicines to treat stroke, and missed days of work.



What health issues or lifestyle traits make patients more or less likely to have a stroke?
Can we use specific features to make reliable prediction of a stroke?

Stakeholders

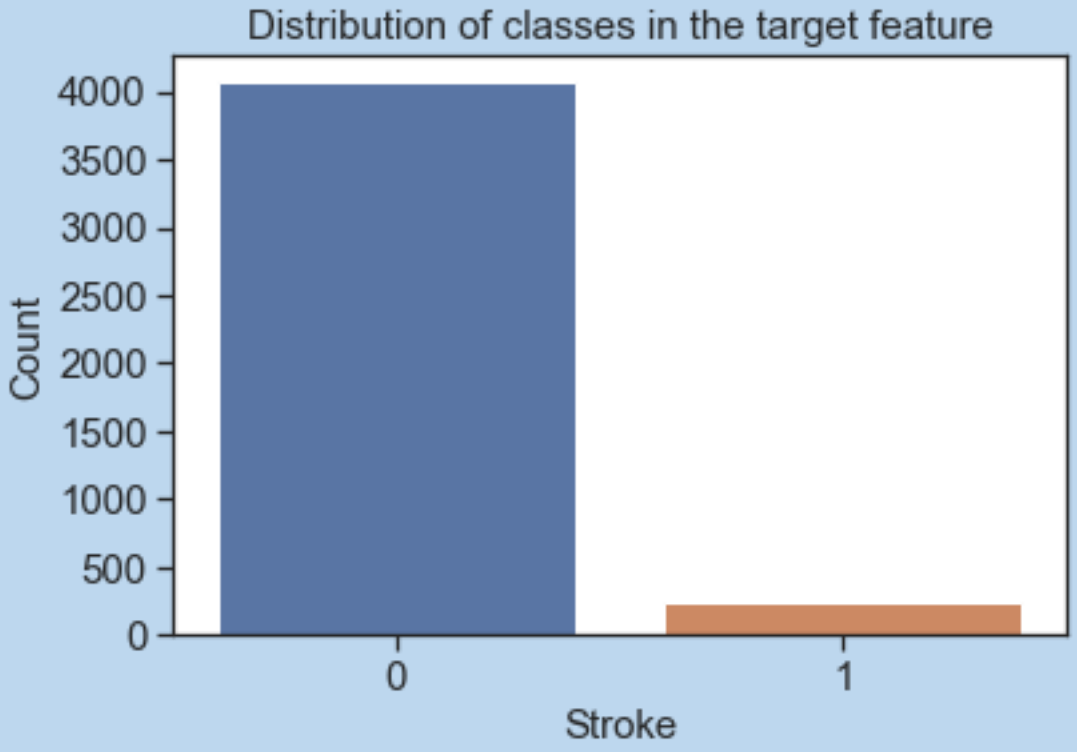
- Health Care Providers
- Health Insurance Companies
- Medical Professionals
- Patients that may be affected by a stroke as well as their family members

The Data

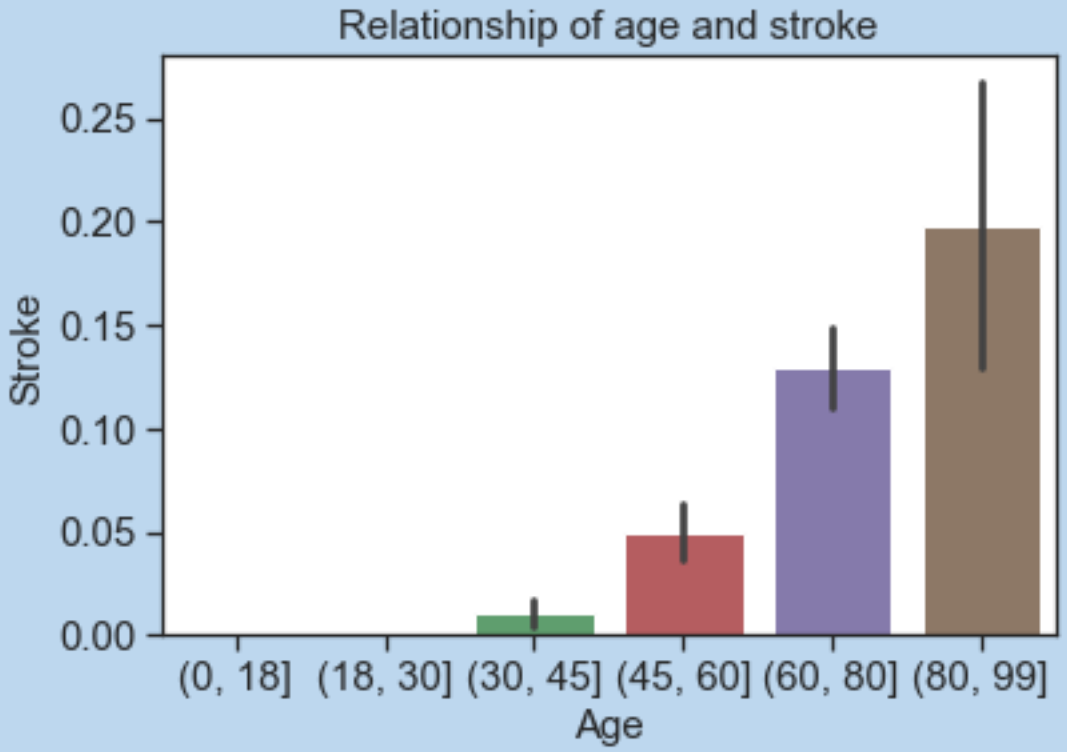
- The data was acquired from kaggle.com
- Number of features: 12
- Number of rows: 5110
- File format: .csv
- Each row represents a unique patient
- ‘id’ column was dropped as it contained irrelevant information.
- The rows that contained age values below 17 were dropped.

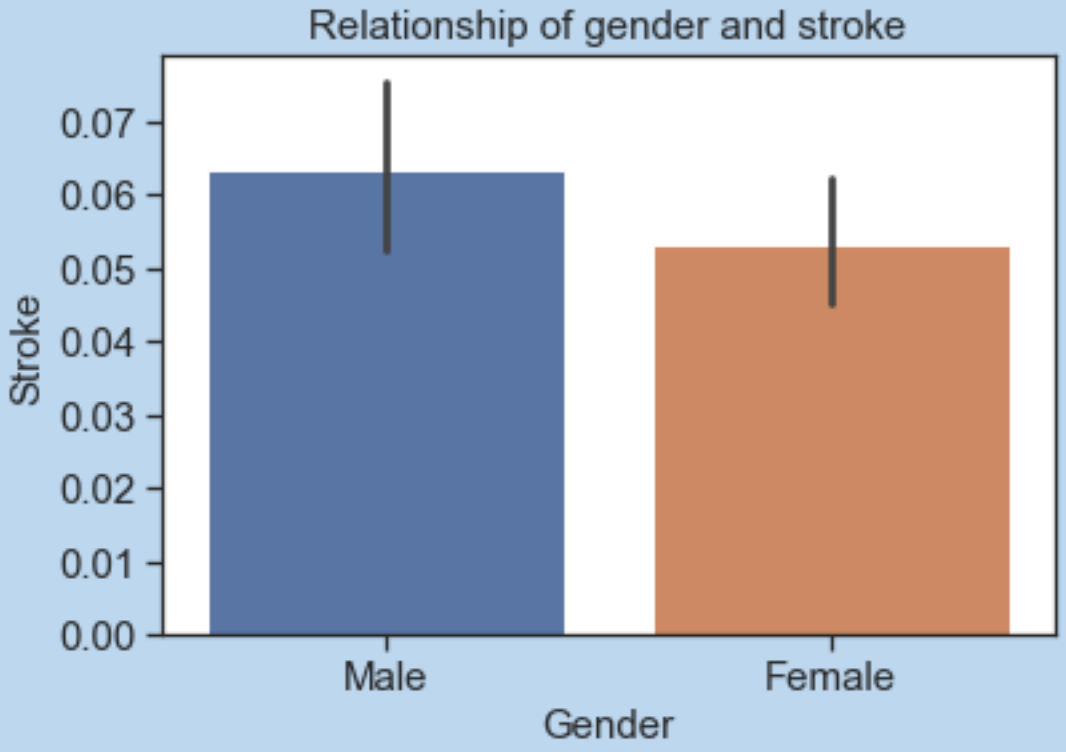
Feature	Datatype	Description
id	int64	Unique ID
gender	object	Male/ Female
age	float64	Applicant age
hypertension	int64	0- If no hypertension, 1- If hypertension indicated
heart_disease	int64	0- If no heart disease, 1- If heart disease indicated
ever_married	object	Yes/No
work_type	object	Government job/ Self-employed/ Private/ Children
residence_type	object	Rural/ Urban
avg_glucose_level	float64	Number indicating average glucose level
bmi	float64	Number indicating BMI score
smoking_status	object	Formerly smoked, Never smoked, Smokes, Unknown
stroke	object	0- If no stroke 1- If stroke indicated

Exploratory Data Analysis
Target Feature

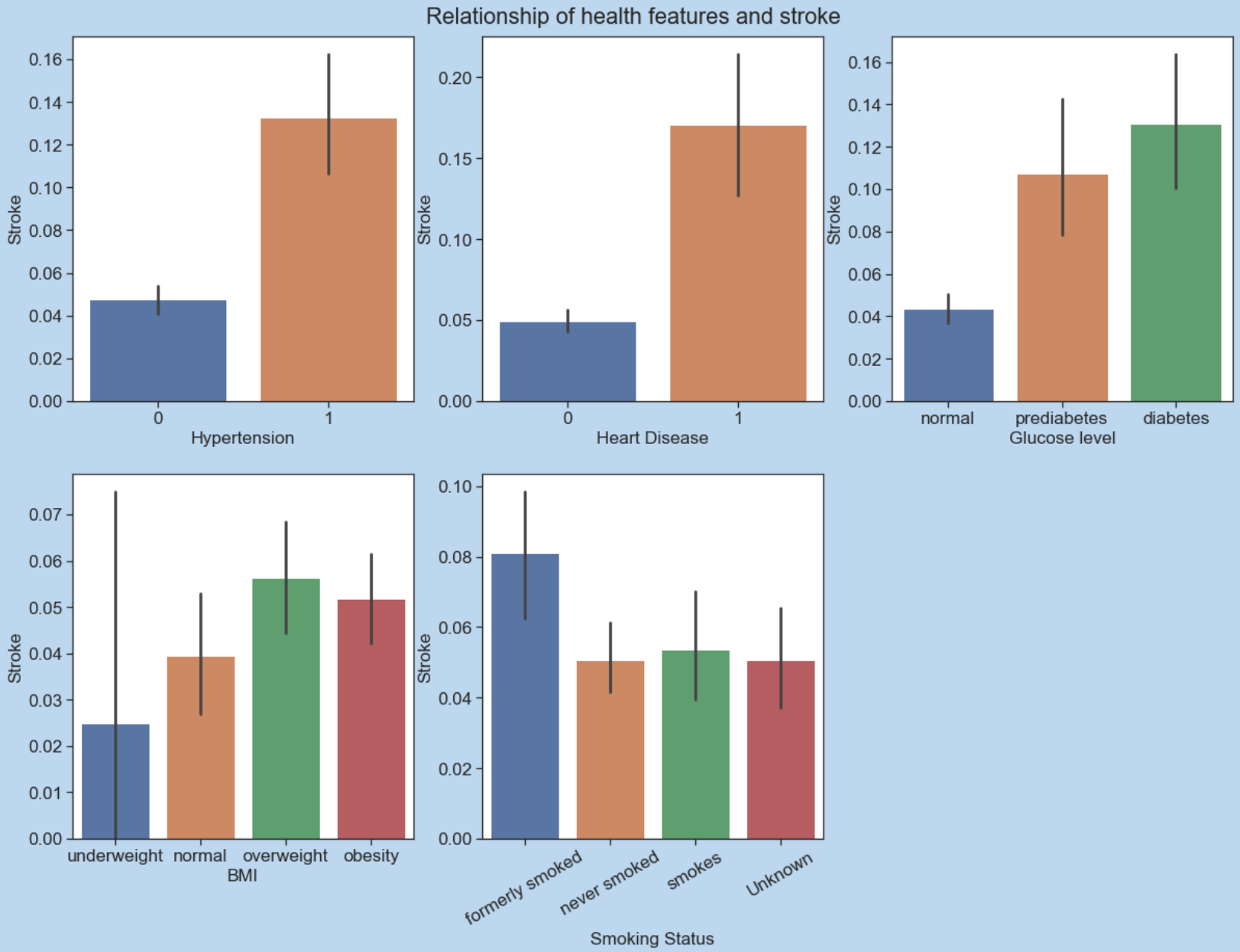


Exploratory Data Analysis
Demographic Features

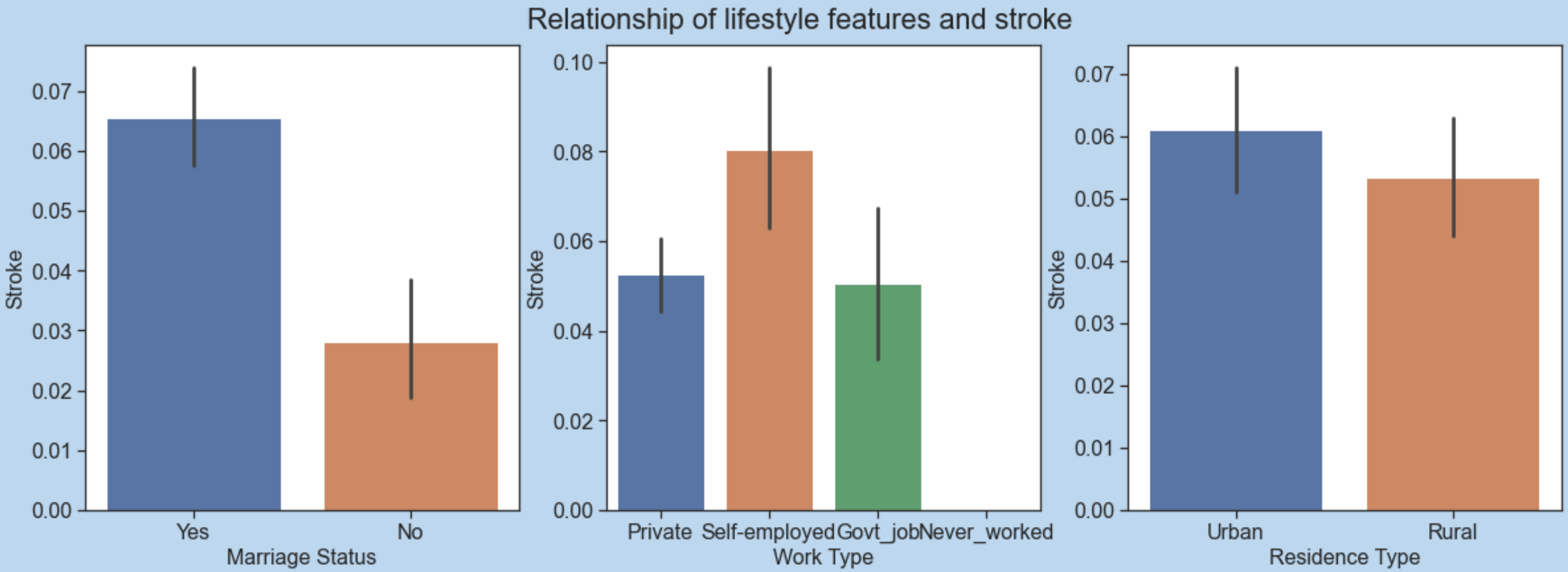




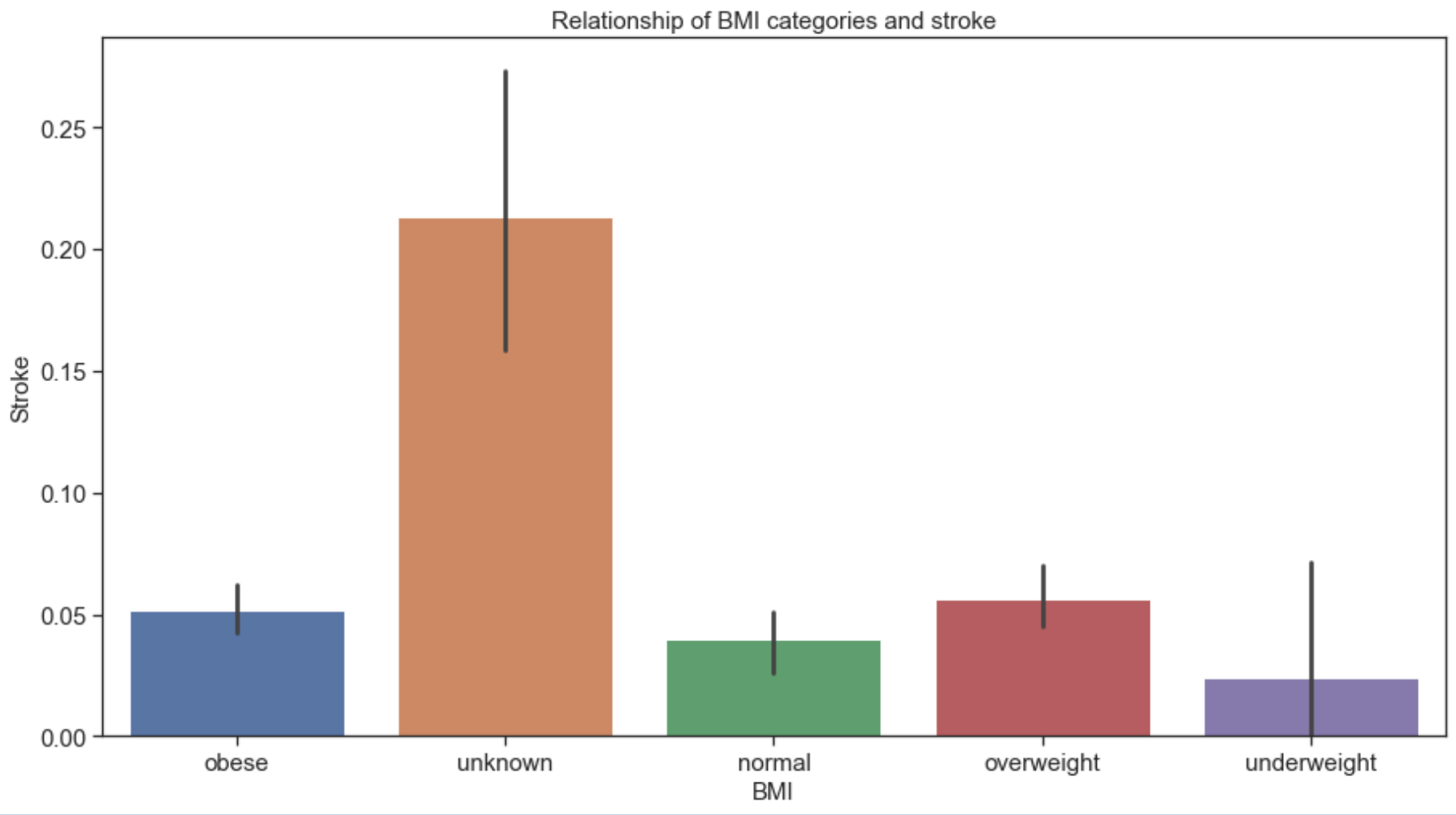
Exploratory Data Analysis
Health Features



Exploratory Data Analysis
Lifestyle features

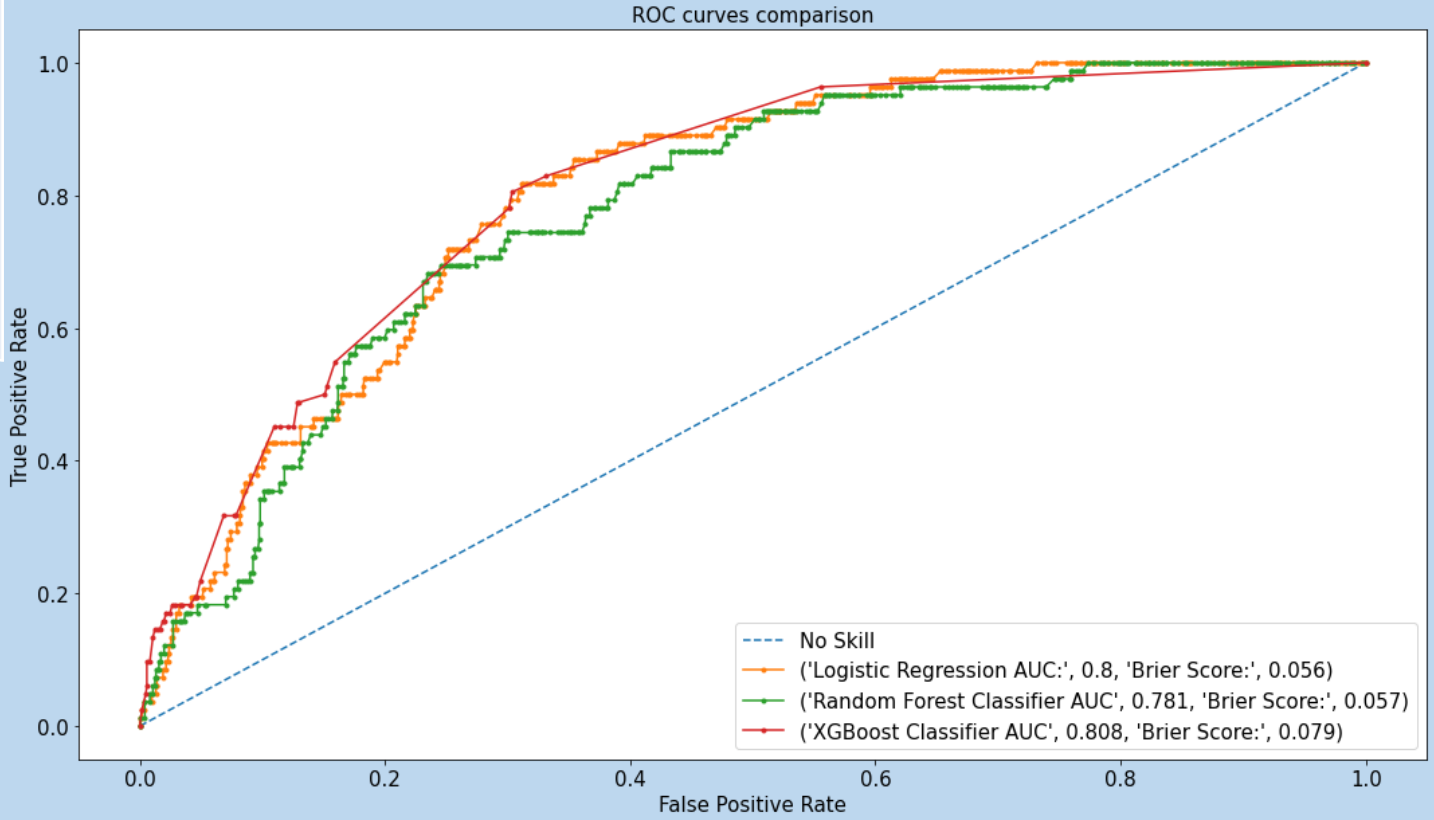


Pre-processing

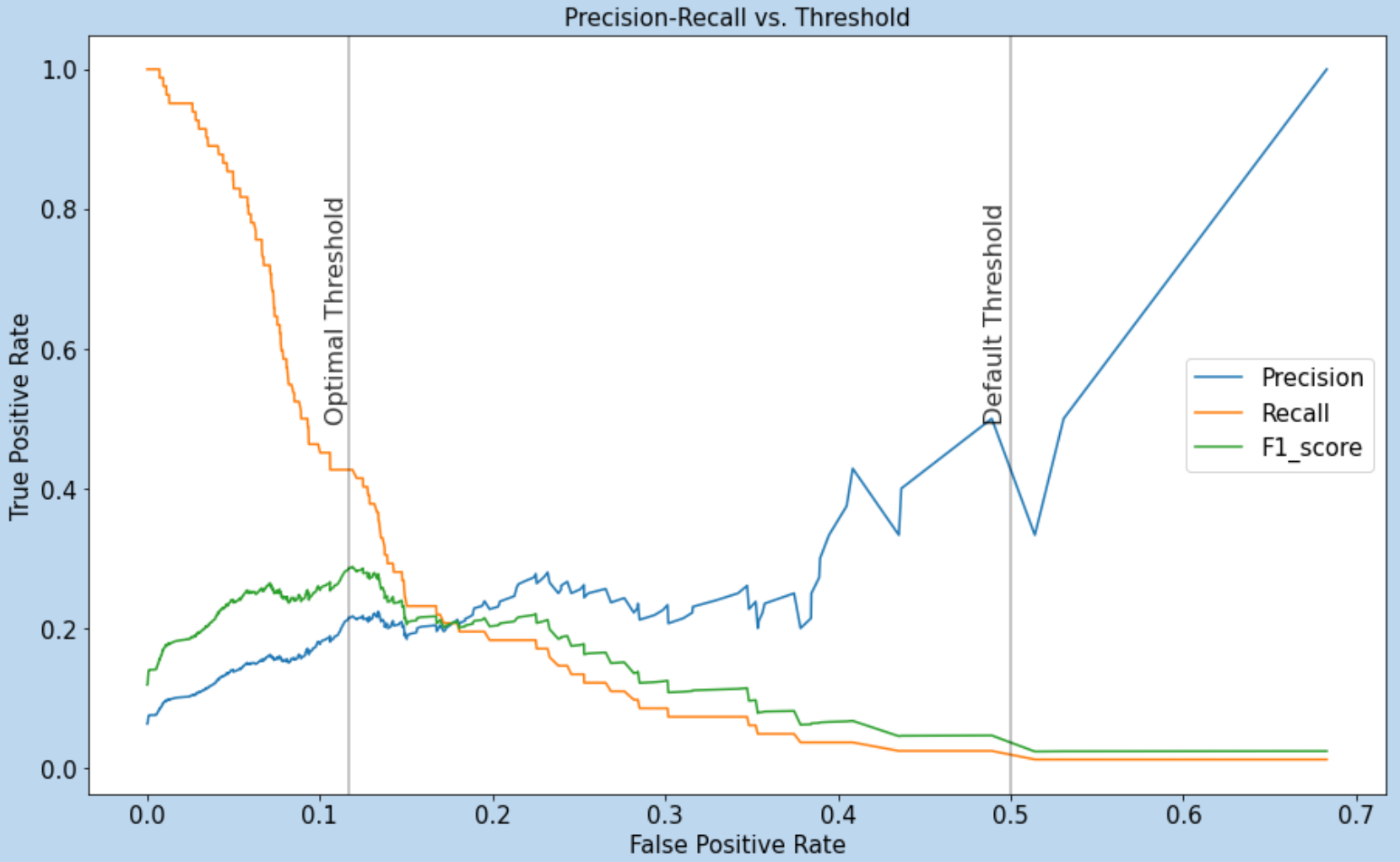


Modeling

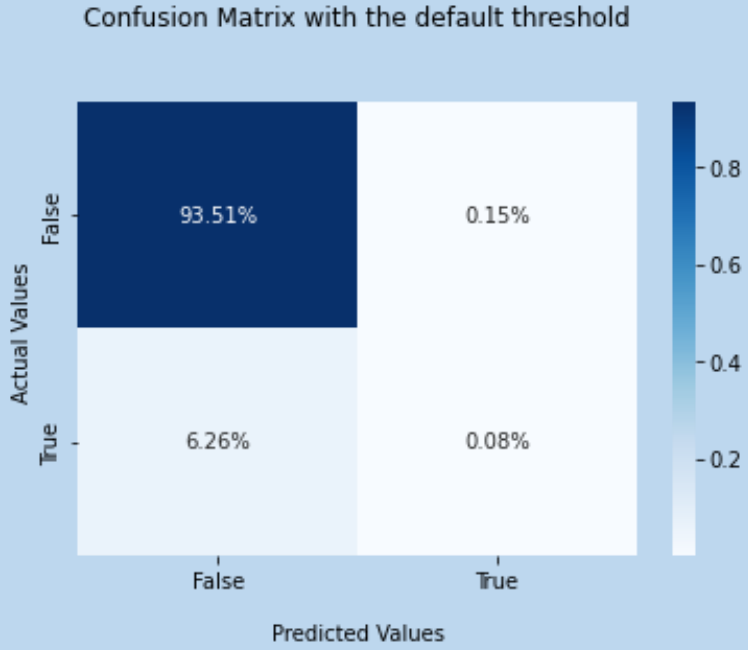
Model	Hyperparameters	AUC score	Brier score
Logistic Regression	C = 10, penalty = elasticnet, solver = saga	0.8	0.056
Random Forest Classifier	criterion = entropy, max_depth = 10, max_features = sqrt, n_estimators = 1000	0.7805	0.057
XGBoost Classifier	booster = dart, eta = 0.001, num_round = 19, gamma = 0.2, max_depth = 4, reg_alpha = 0.001, n_estimators = 1000, min_child_weight = 2	0.808	0.079



Modeling

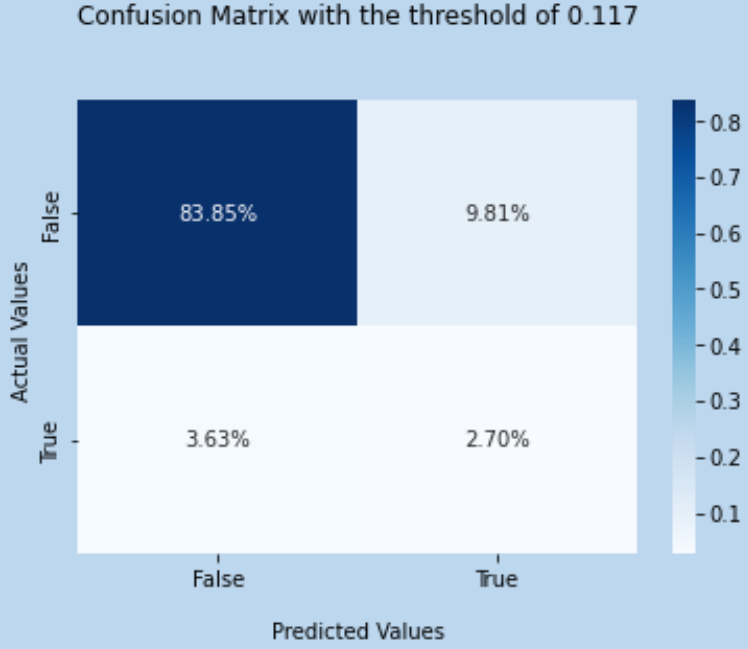

Precision-Recall vs. Threshold
True Positive Rate
False Positive Rate
Optimal Threshold
Default Threshold
Precision
Recall
F1_score
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Modeling



Classification report with the default threshold

	precision	recall	f1-score	support
0	0.94	1.00	0.97	1212
1	0.33	0.01	0.02	82
accuracy			0.94	1294
macro avg	0.64	0.51	0.50	1294
weighted avg	0.90	0.94	0.91	1294



Classification report with optimal threshold of 0.117:

	precision	recall	f1-score	support
0	0.96	0.90	0.93	1212
1	0.22	0.43	0.29	82
accuracy			0.87	1294
macro avg	0.59	0.66	0.61	1294
weighted avg	0.91	0.87	0.89	1294

Conclusion

- Credibility of a model predictions should be taken with a grain of salt
- Other factors, that can increase a likelihood of a stroke, were not studied in this project.