Capstone Final Report

1. Problem Statement

Is it possible to predict whether or not an individual will suffer a stroke based on input parameters such as gender, age, various diseases, smoking status, etc.? What factors affect stroke occurrence and can we predict the likelihood of a stroke?

2. Context and Goals

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. 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.

The goal of this project would be to develop the best model for predicting whether or not an individual had a stroke and being able to accurately predict the likelihood of a stroke based on input parameters such as age, gender, etc.

3. Dataset

For this project we're using a stroke prediction dataset from Kaggle.

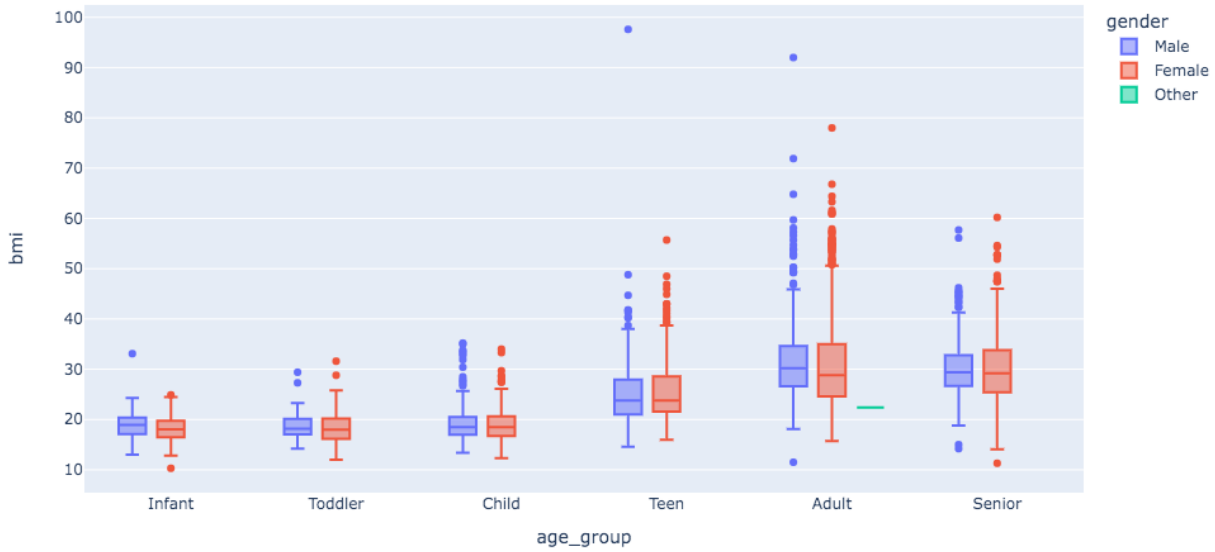
The feature list and the target variable:

1. id: unique identifier
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6. ever_married: "No" or "Yes"
7. work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8. Residence_type: "Rural" or "Urban"
9. avg_glucose_level: average glucose level in blood
10. bmi: body mass index
11. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12. stroke: 1 if the patient had a stroke or 0 if not

Note: "Unknown" in smoking_status means that the information is unavailable for this patient.

4. Data Cleaning and Data Wrangling

First, we identified that the BMI column had 4% missing values. The solution was to impute the missing BMI values based on the median BMI value for different age groupings.

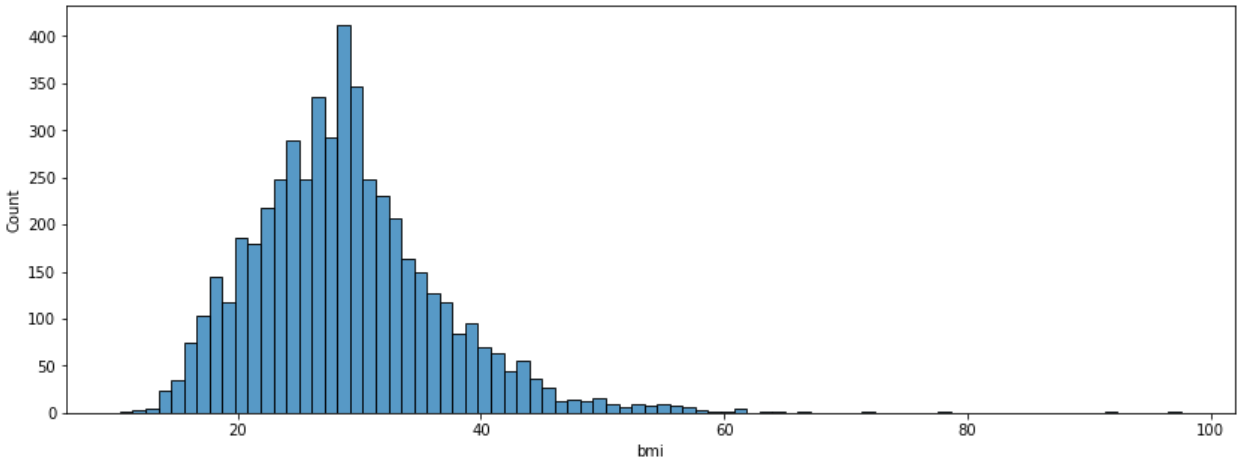


There were no duplicate values for the ID column in this dataset, all ID values were unique. Data types were also appropriate for each column.

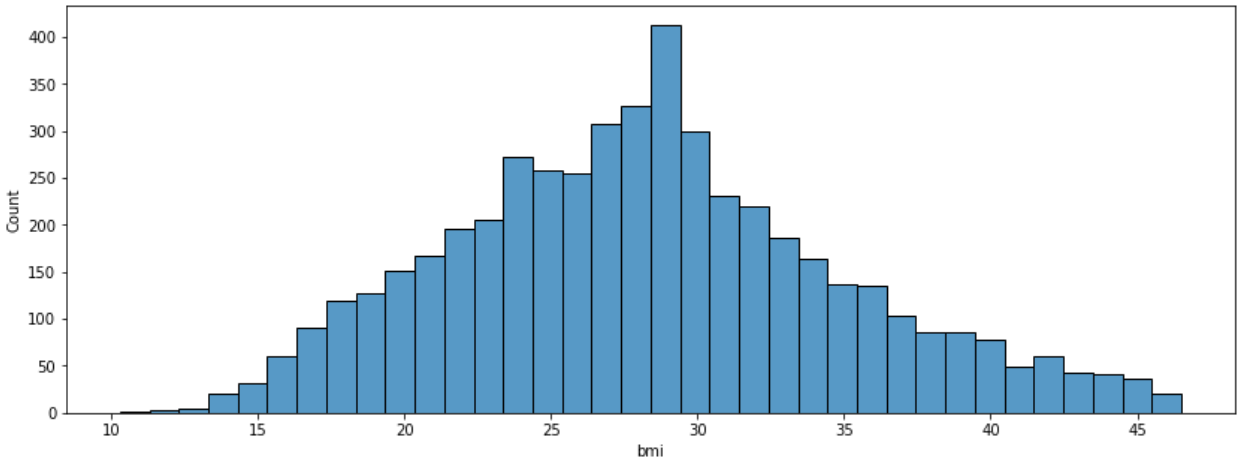
One of the issues was the “Other” value in the “Gender” column. Since there was only one value that had the “Other” category, this row was dropped.

Next we checked for outliers. Age column did not have any outliers. BMI and average glucose level columns had outliers. There were 84 outliers in the average glucose level column who had a stroke, which is a lot of data. Removing them would have led to loss of data. BMI only had 3 outliers who had a stroke, so removing these wouldn’t have affected data as much. We removed the BMI outliers and checked to make sure BMI distribution looked less skewed.

BMI distribution before:

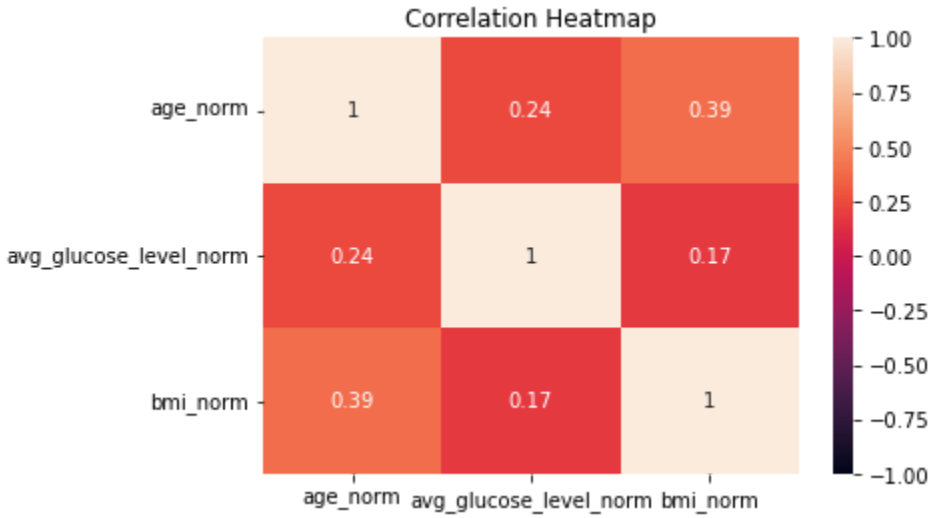


BMI distribution after removing the outliers:

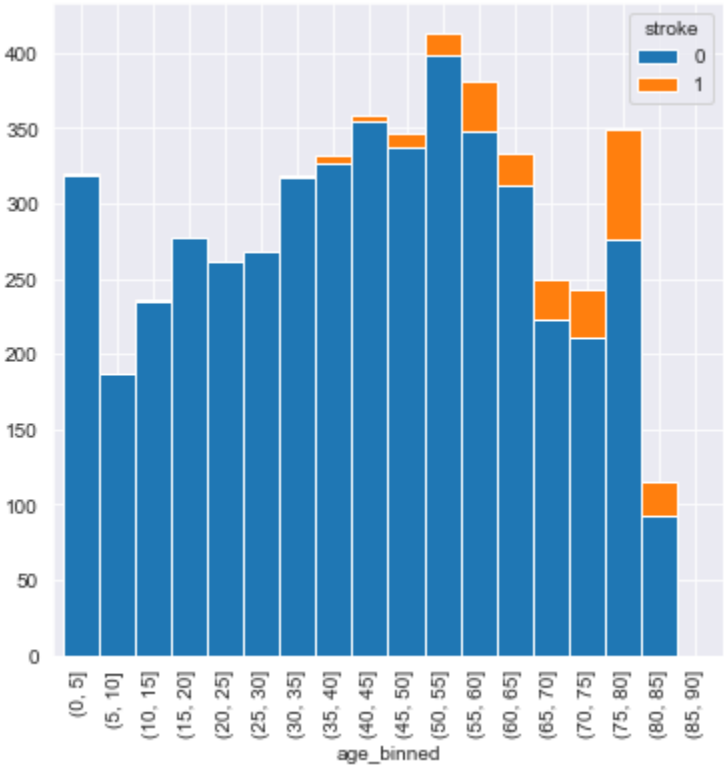


5. Exploratory Data Analysis

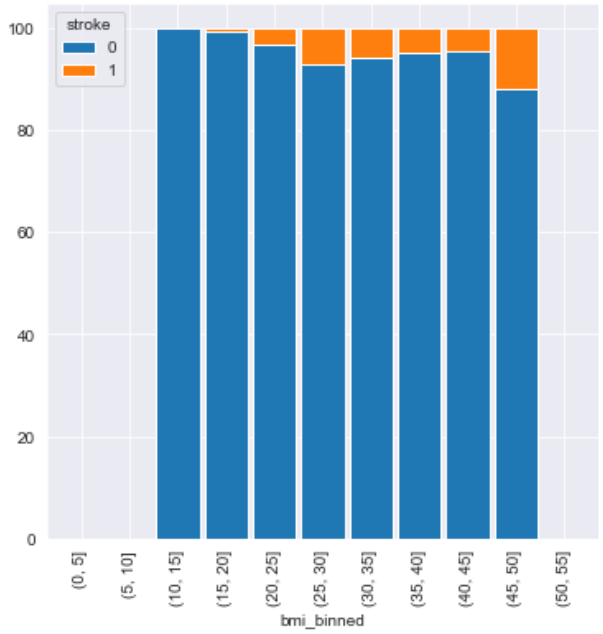
BMI and Age are positively correlated, though the association is not strong:



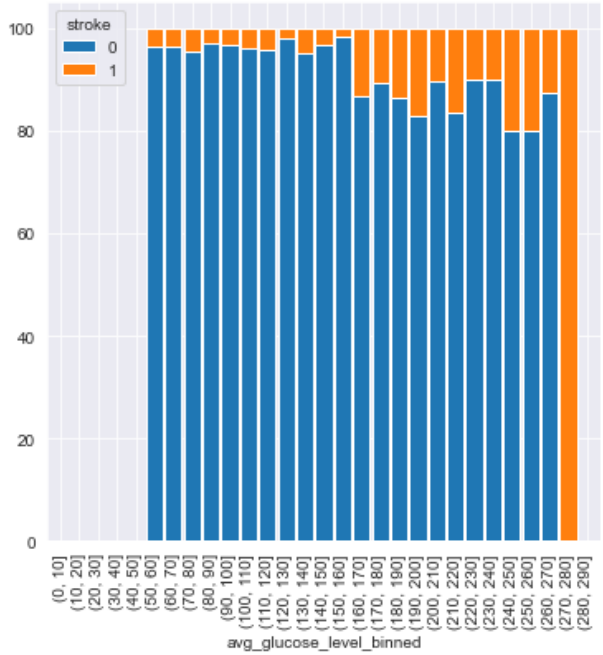
Older patients are more likely to suffer from a stroke:



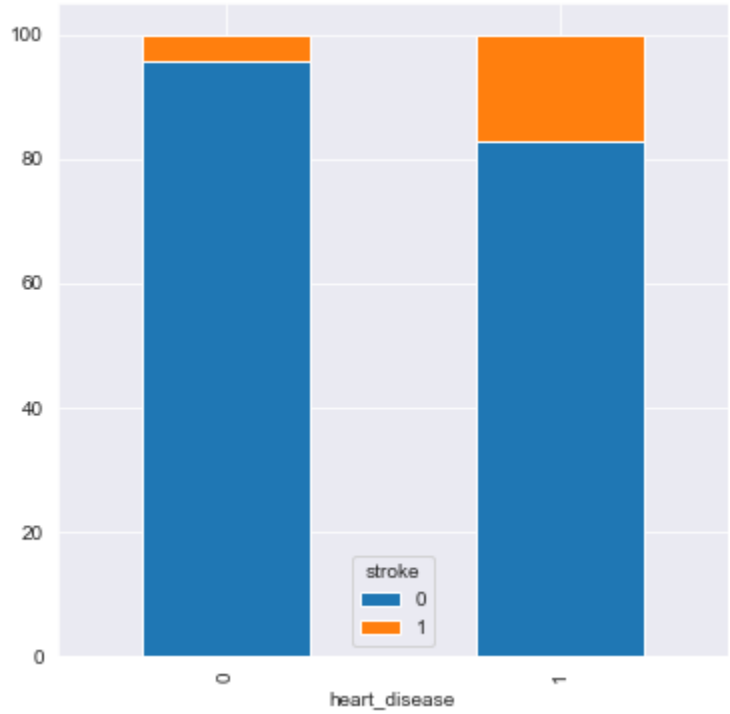
Higher BMI is associated with a higher chance of stroke:



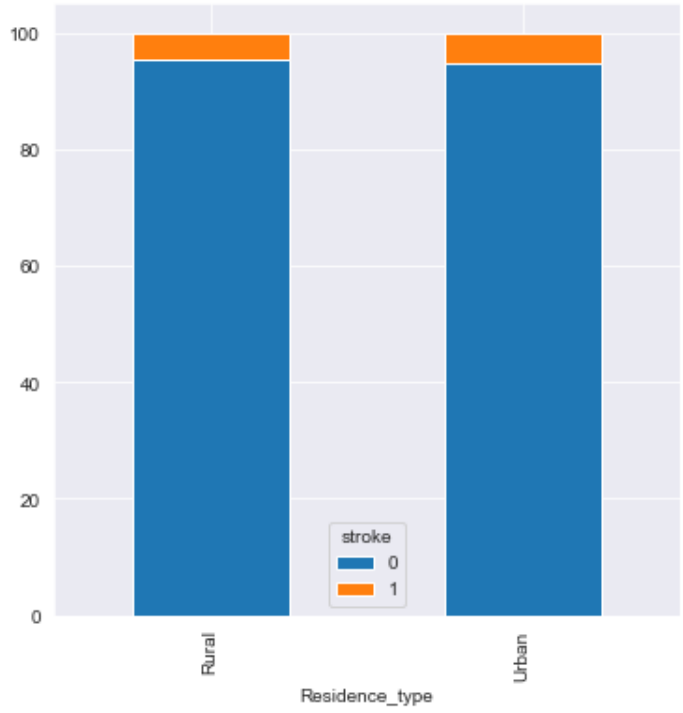
Higher average glucose level is also associated with a higher change of stroke:



Patients with hypertension and heart diseases had higher numbers of strokes compared to patients with no hypertension / heart disease:

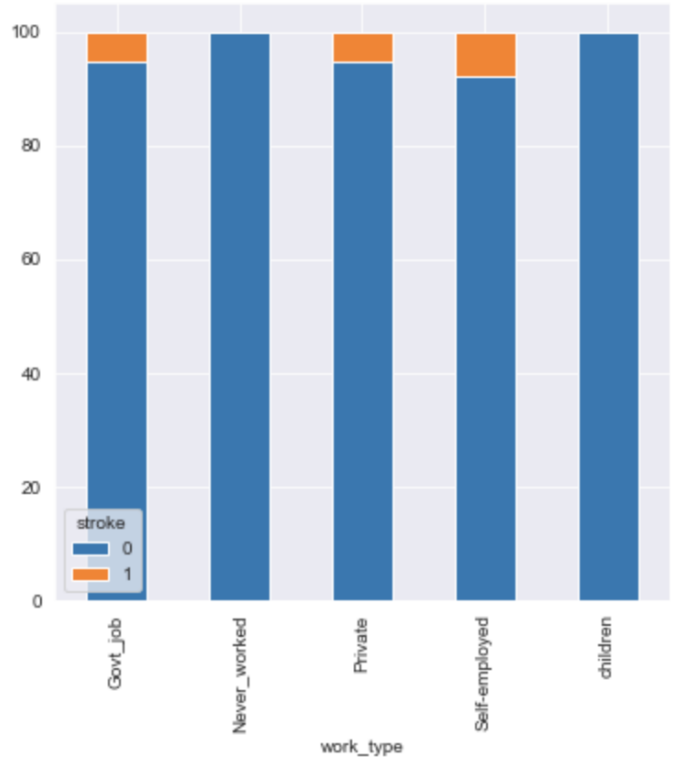


Gender and residence type (urban / rural) didn't seem to be associated with higher likelihood of a stroke:



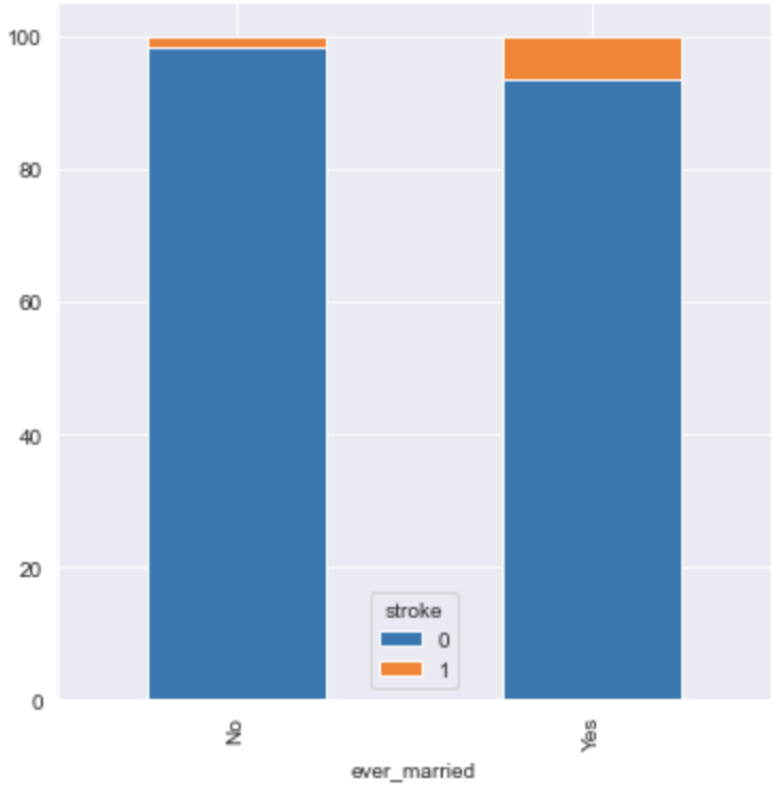
Work type and marital status seem to be associated with higher change of stroke but the categories that had higher number of strokes included people who were older. Patients with work type of (self-employed / government job / private) had higher numbers of strokes but the age was also higher for people in these work sectors (average age for government jobs - 51, average age for self-employed - 60, average age for private - 46):

	age	
	count	mean
work_type		
Govt_job	640	50.871875
Never_worked	22	16.181818
Private	2838	45.629316
Self-employed	801	60.357054
children	687	6.841339



Number of strokes was higher among married people but the mean age for married people is also higher (54 for married vs. 22 for not married):

	age	
	count	mean
ever_married		
No	1730	21.874566
Yes	3258	54.515961



6. Data Preprocessing and Modeling

The stroke prediction dataset is imbalanced, so I applied SMOTE technique to balance out the 2 groups (stroke / no stroke). The data was split into 80% train / 20% test groups. The columns were standardized using StandardScaler.

3 models were tested: Logistic Regression, Support Vector Machine (SVM), and Random Forest classifier.

Logistic Regression

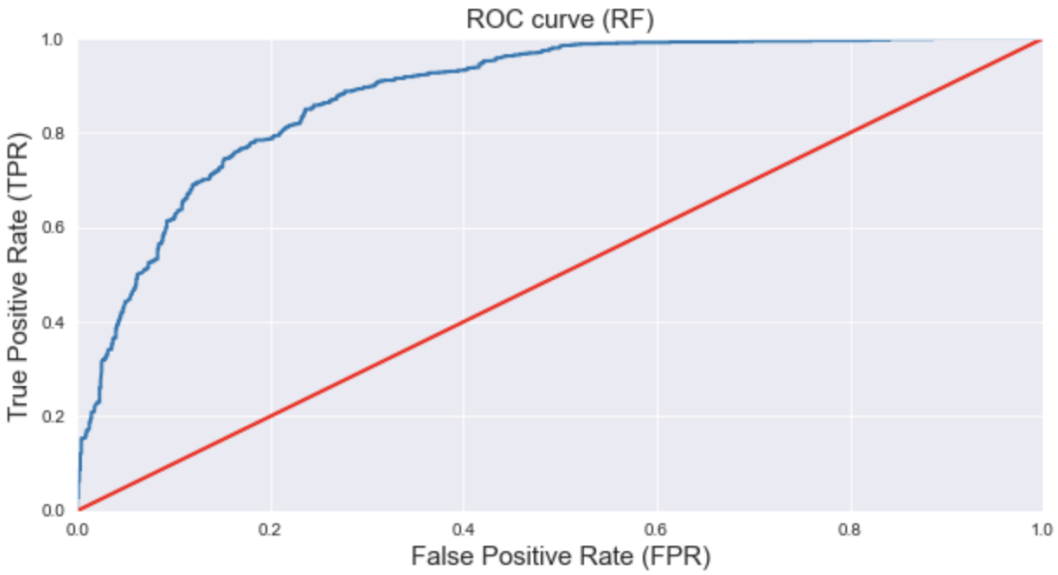
The Logistic Regression model showed **79% accuracy** (TN + TP / (TP + FP + TN + FN). The model was run using RandomizedSearchCV (see metrics file for parameters). The ROC curve:

Support Vector Machine (SVM)

The SVM model showed 93% accuracy (see metrics file for parameters). ROC curve:

Random Forest

Random Forest model gave an accuracy score of 80% (see metrics file for parameters). ROC curve:



7. Model Comparison

The **SVM model** gave the best accuracy and recall. We care about recall (TP / TP + FN) because we don't want to misdiagnose people who are likely to suffer from stroke as negative. Comparison of the models:

	Algorithm	ROC-AUC train score	ROC-AUC test score
0	Logistic Regression	0.888266	0.881759
1	SVM	0.957949	0.906161
2	Random Forest	0.894338	0.878049