Predicting stroke probability

T5 BOOTCAMP DATA SCIENCE PROJECT

Outlines

- Project goal
- Dataset
- Data cleaning
- Data processing
- Visualize data
- Models

Project Goal

The goal of this project is to use classification model to predict which patients are close to have stroke and compute the accuracy of this model. These outcomes will help to focus on potential patients and avoid them being in stroke critical case.

Dataset

- Data provided by Kaggle.com has been used in this project.
- The dataset contains 5110 unique records with 12 attributes for each, collecting from 2995 females and 2115 males. The last column contains ‘1’ if the patient had stroke and ‘0’ if he or she hadn’t. A few attributes like age, marriage, and hypertension will be consider in prediction model. The dataset divided in three parts:

- 20% for testing.
- 60% for training.
- 20% for validation.

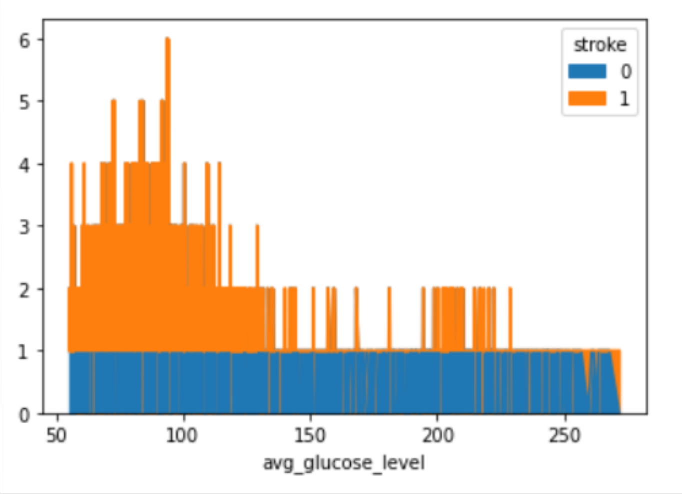
Data Cleaning

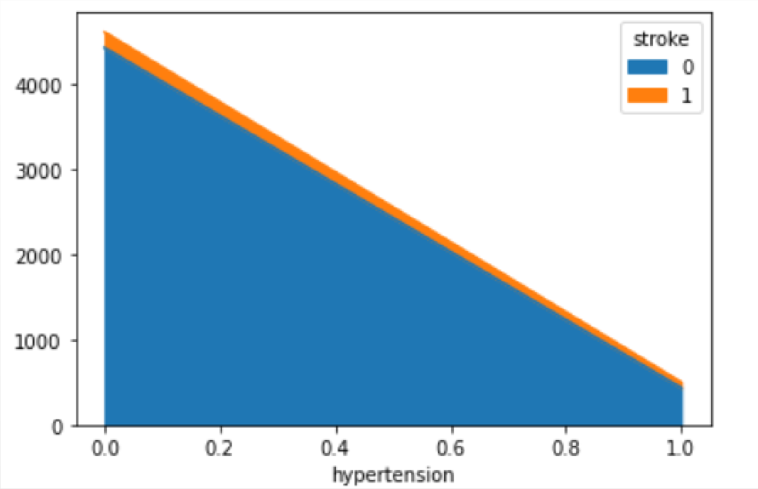
- Missing Data
  - Replaced with mean value
- Irregular Data (Outliers)
  - Replaced with mode value
- Unnecessary Data
  - Deleted it
- Inconsistent Data
  - Nothing

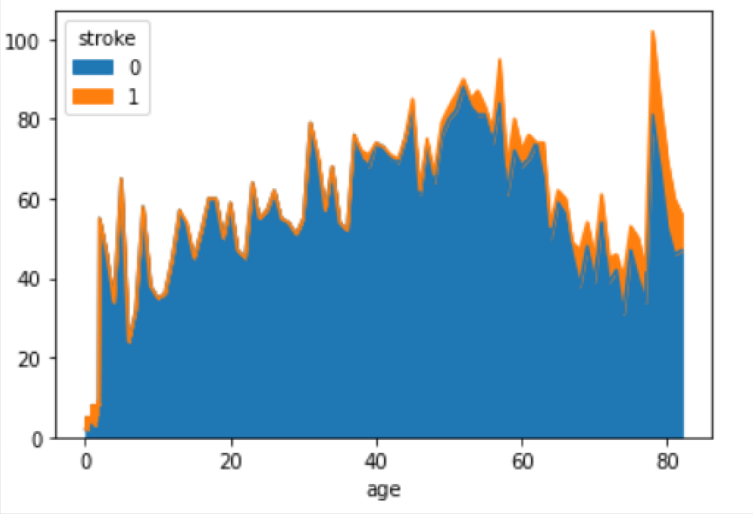
Data Processing

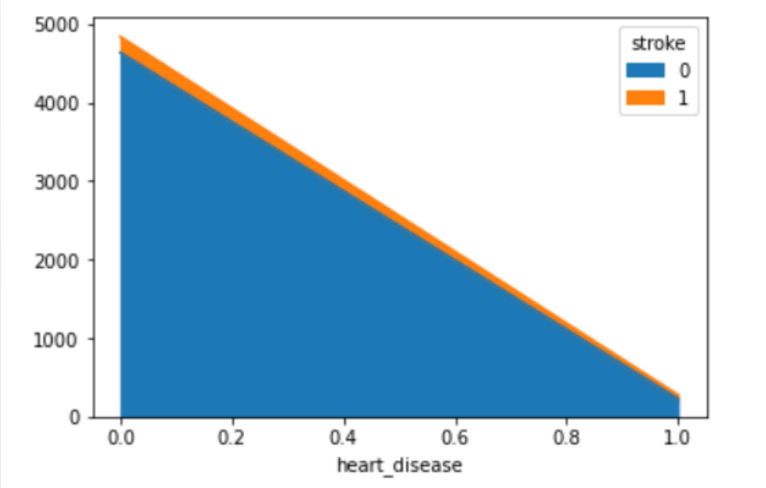
- Convert categorical variables to numbers by LabelEncoder from sklearn.
- Scale values by using StandardScaler from sklearn.
- Use SMOTE from imblearn to solve the imbalence data

Data Visualization

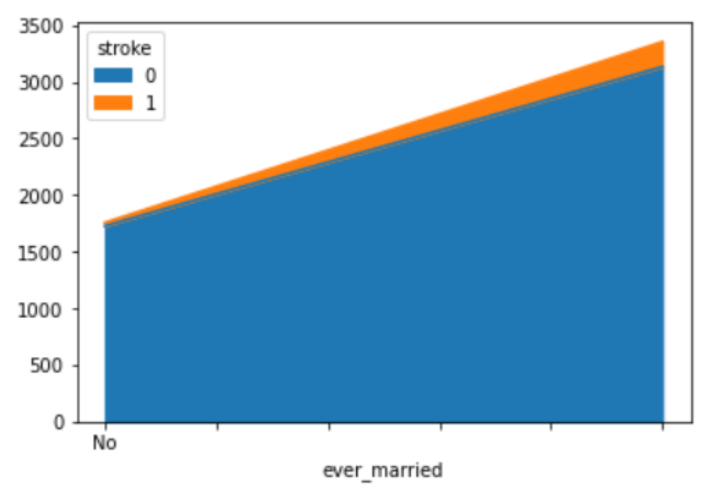


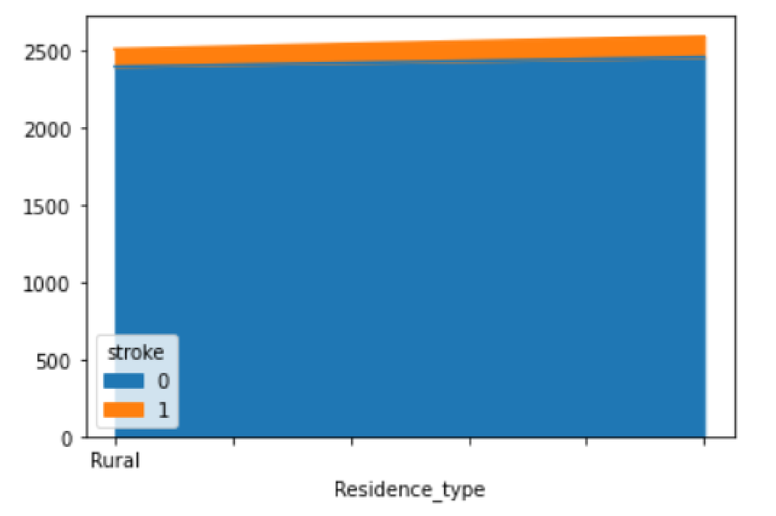


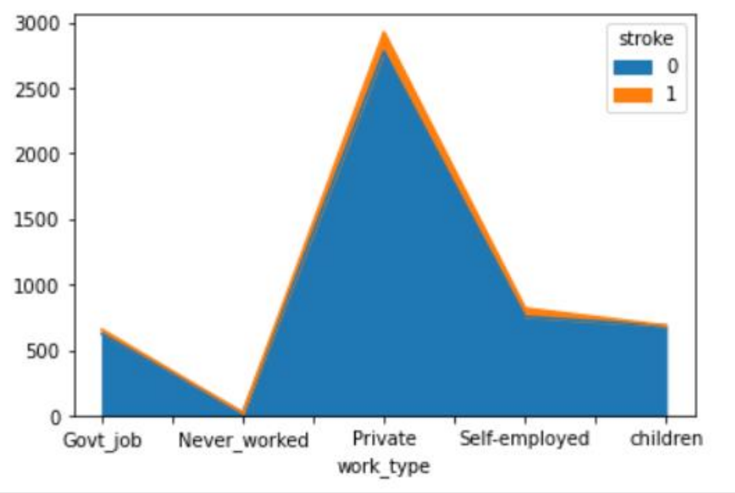


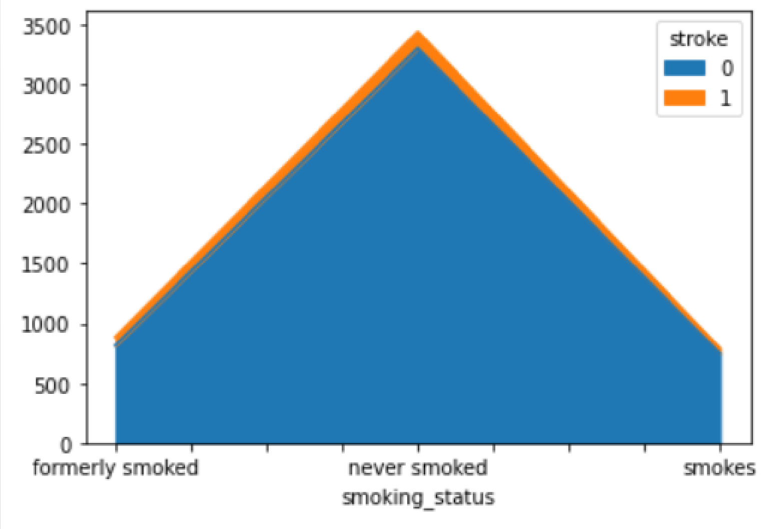


Data Visualization

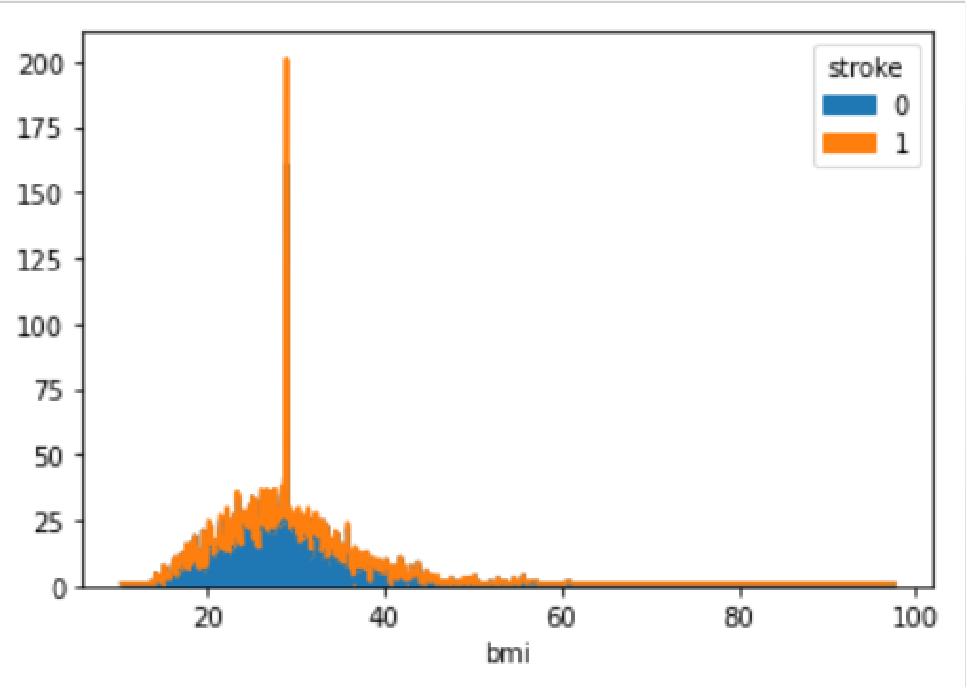


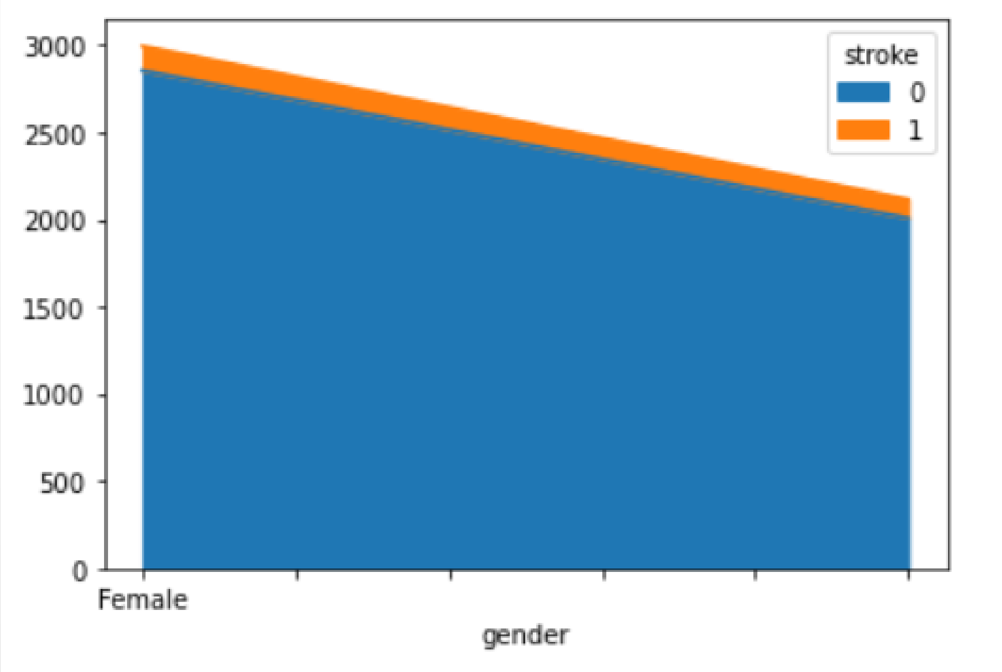






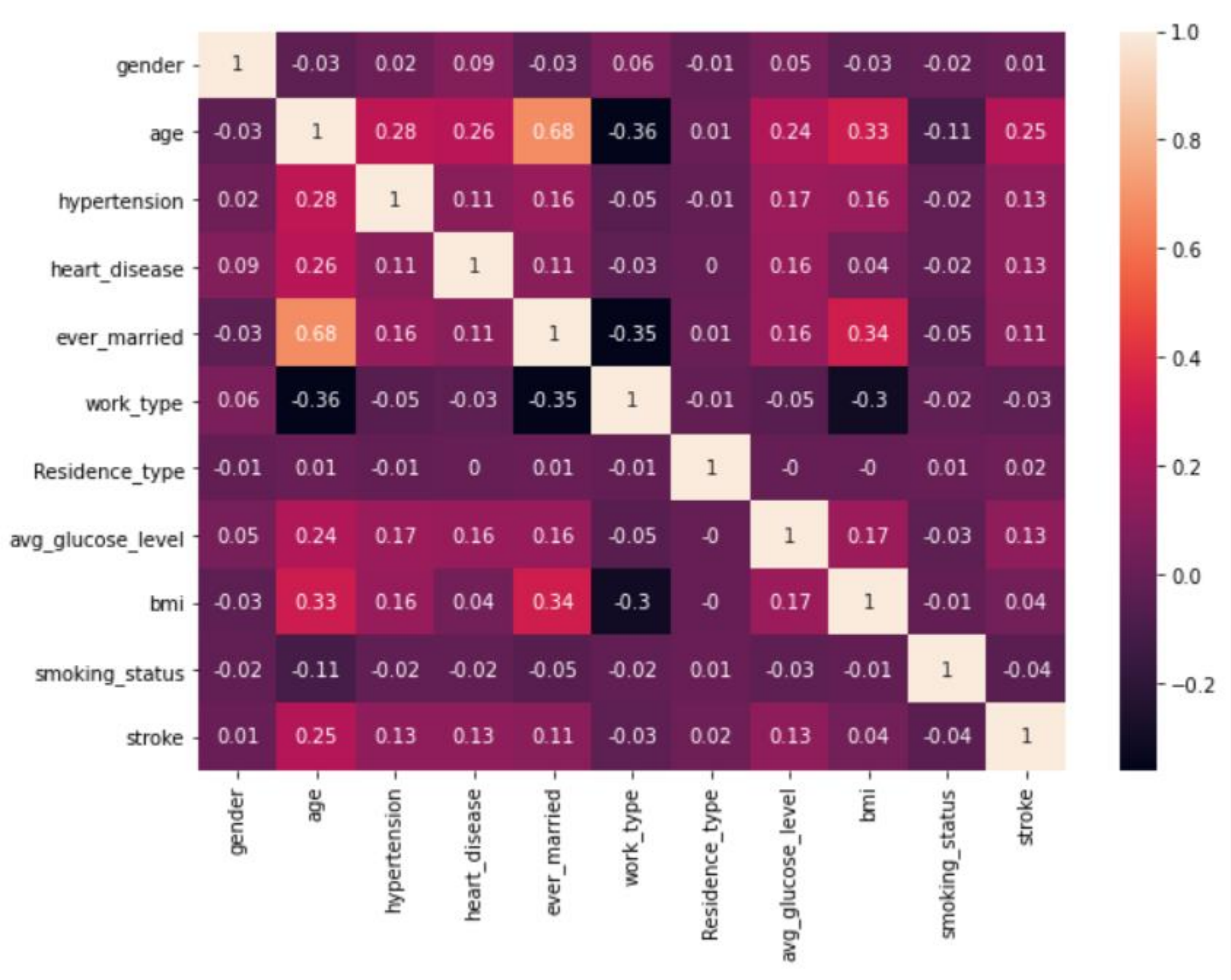
Data Visualization





Data Visualiza

All together



	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
gender	1	-0.03	0.02	0.09	-0.03	0.06	-0.01	0.05	-0.03	-0.02	0.01
age	-0.03	1	0.28	0.26	0.68	-0.36	0.01	0.24	0.33	-0.11	0.25
hypertension	0.02	0.28	1	0.11	0.16	-0.05	-0.01	0.17	0.16	-0.02	0.13
heart_disease	0.09	0.26	0.11	1	0.11	-0.03	0	0.16	0.04	-0.02	0.13
ever_married	-0.03	0.68	0.16	0.11	1	-0.35	0.01	0.16	0.34	-0.05	0.11
work_type	0.06	-0.36	-0.05	-0.03	-0.35	1	-0.01	-0.05	-0.3	-0.02	-0.03
Residence_type	-0.01	0.01	-0.01	0	0.01	-0.01	1	-0	-0	0.01	0.02
avg_glucose_level	0.05	0.24	0.17	0.16	0.16	-0.05	-0	1	0.17	-0.03	0.13
bmi	-0.03	0.33	0.16	0.04	0.34	-0.3	-0	0.17	1	-0.01	0.04
smoking_status	-0.02	-0.11	-0.02	-0.02	-0.05	-0.02	0.01	-0.03	-0.01	1	-0.04
stroke	0.01	0.25	0.13	0.13	0.11	-0.03	0.02	0.13	0.04	-0.04	1

Models

- Models' accuracy
- Logistic Regression: 0.786692759295499
- Random Forest: 0.910958904109589
- Decision Tree: 0.8688845401174168
- KNN: 0.8140900195694716
- Naive Bayes: 0.8140900195694716
- KMeans: 0.8551859099804305

Thank You