Prediction of Stroke

Yiwei Sang
Data Science Initiative
Brown University

Oct 22, 2019

https://github.com/yiweisang97/data1030_project.git

Introduction

- Predicting whether an individual is likely to have stroke
- Stroke is the No. 5 cause of death and a leading cause of disability in the United States. Machine learning projects
- Classification
- https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data

Data Preprocessing

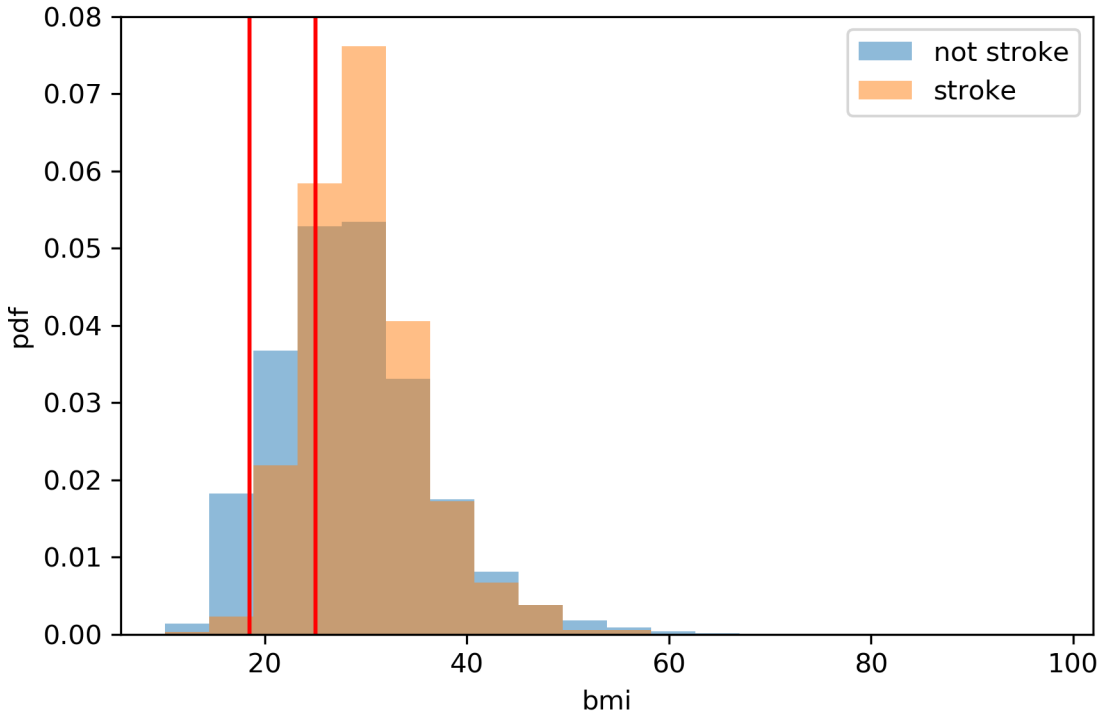
- Original dataset:
  - 43400 observations
  - 11 features:
    - Numerical: id, age, avg_glucose_level, bmi
    - Categorical: gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status
  - Target variable - stroke:
    - 0: 98.195853%
    - 1: 1.804147%
  - Missing value:
    - smoking_status: 30.63%
    - bmi: 3.37%

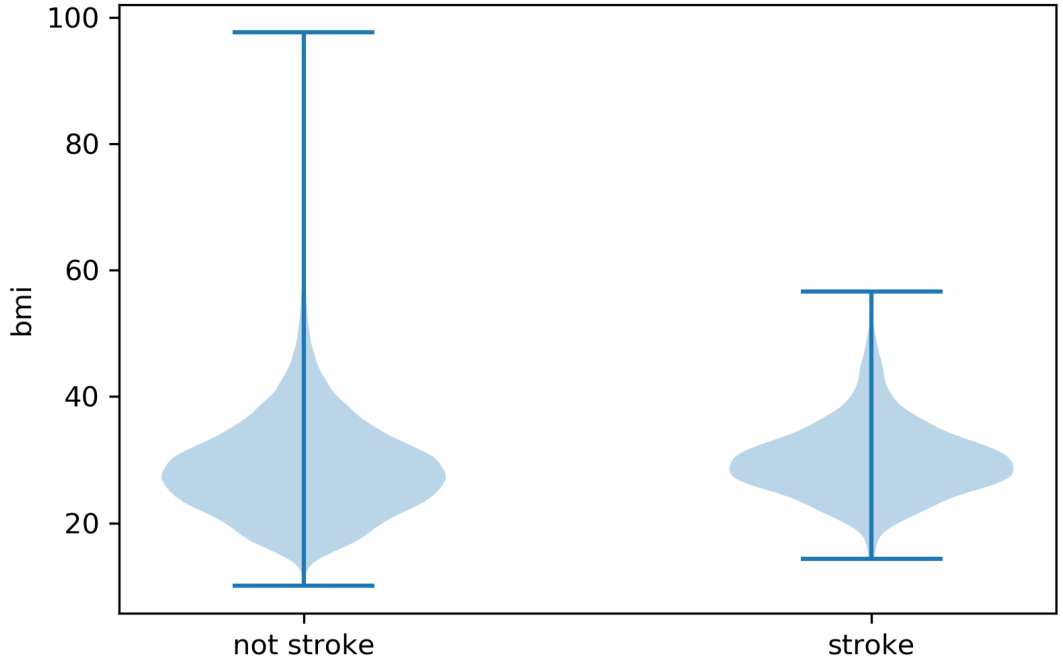
- Preprocessing:
  - Drop column id
  - Missing value:
    - smoking_status: replace with ‘missing’
    - bmi: average of five RandomForest Imputer
  - MinMaxScaler & OneHotEncoder
  - No need to preprocess hypertension and heart_disease and target variable

- After preprocessing:
  - 43400 observations
  - 21 features

Exploratory Data Analysis

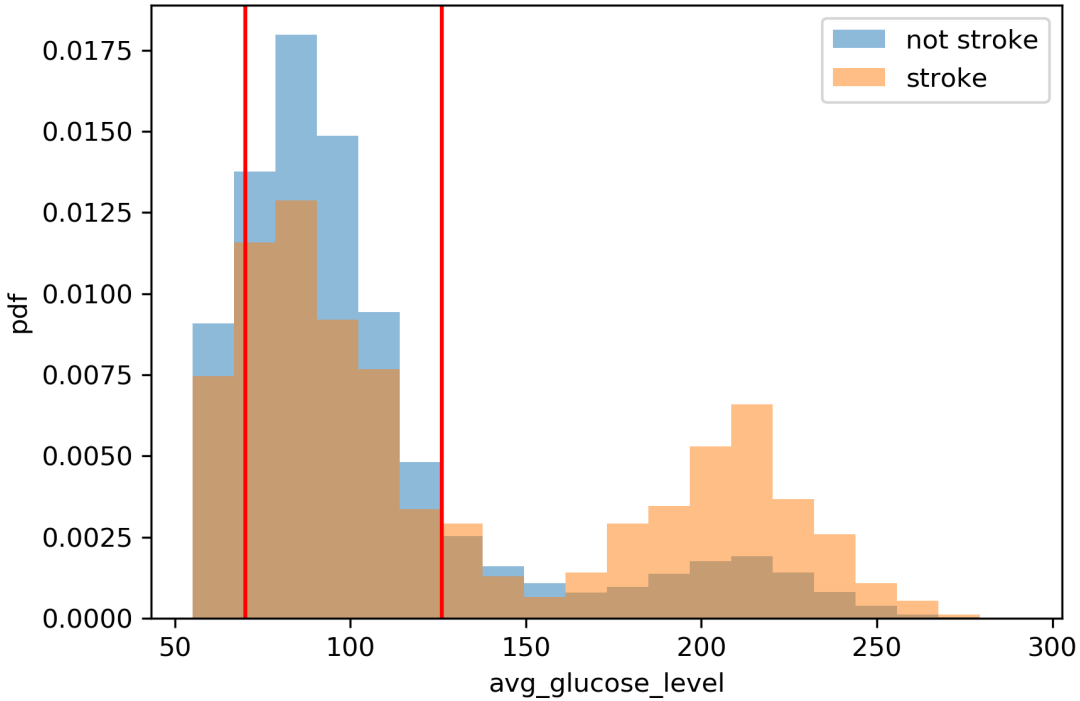
bmi

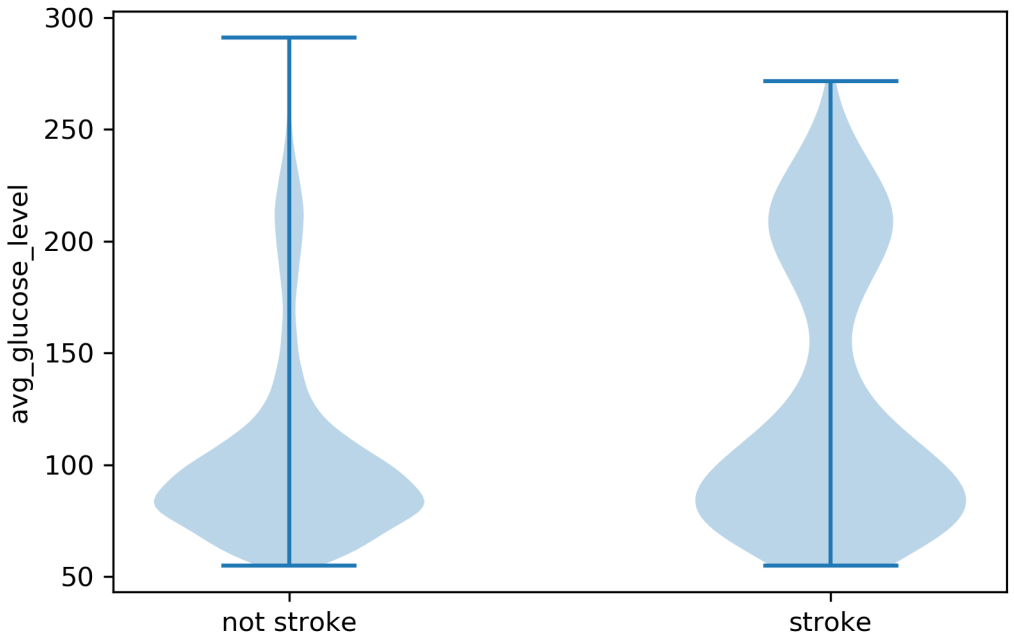




Exploratory Data Analysis

avg_glucose_level





Exploratory Data Analysis

smoking_status