Data1030 Project Proposal

Yiwei Sang

Introduction

Stroke is a serious illness. It is No.5 cause of death and leading cause of disability in the United States. It would be nice if we can predict whether an individual is likely to have a stroke and thus prevent it.

In this project, I will construct a binary classification model that predict whether an individual has a stroke or not. The target variable(stroke) is 1 if the individual had a stroke before, 0 if the individual had not. Hopefully, the model can be generalized to predict the tendency of the general public to have a stroke.

Dataset

Link to dataset: https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data

The data is obtained from Kaggle. The original data includes two datasets: train_2v.csv and test_2v.csv. The former dataset includes the target variable stroke while the latter does not. Since classification is supervised machine learning, target variable is needed. Therefore, I will use train_2v.csv for training and testing. I renamed train_2v.csv to stroke.csv. The dataset has 43400 entries and 12 columns. The first column is 'id', which identifies unique individuals; the last column is the target variable 'stroke' and the rest ten columns represent ten features:

- gender: Male, Female
- age: age of patient in years
- hypertension: 1 if the patient suffers from hypertension, 0 if the patient does not
- heart_disease: 1 if the patient suffers from heart disease, 0 if the patient does not
- ever_married: Private, Self-employed, Govt-job, children, Never_worked
- avg_glucose_level: Average glucose level measured after meal, measured in mg/dL
- bmi: Body mass index, measured in kg/㎡.
- smoking_status: never smoked, formly smoked, smokes

The dataset contains NaN values in 'bmi' and 'smoking_status' columns.

There are some public projects that use this dataset:

- There is one classification model that applies tree method.
- There is another classification model that applies logistic regression. This project uses mean of bmi to fill out missing values in 'bmi' column and construct two models, one using entries with available 'smoking_status' and the other sing entries with missing 'smoking_status'.

Preprocessing

First I dropped NaN values in the data. The data now have 29072 entries.
I apply OneHotEncoder to ‘gender’, ‘hypertension’, ‘heart_disease’, ‘ever_married’, ‘work_type’, ‘residence_type’ and ‘smoking status’, since these features are categorical and not ordinal.
I apply MinMaxScaler to ‘age’, ‘avg_flucose_level’ and ‘bmi’, since these features are numerical and are bounded.
There are 22 features in the preprocessed data.

Github repo

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