CS6140 Course Project Proposal
Group 14
Group member: Xichen Liu, Duo Zhao, Keming Xu, Wenshu Yang, Rodolph Mesadieu

Prediction on Stroke based on various models

Basic Idea

According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. We are curious to know if stroke is predictable and what factors may lead to stroke. In this project, we will be building a couple of machine learning models to make predictions on whether a patient is likely to get a stroke, using the dataset of a number of patients' body features and measurements, like gender, age, various diseases, and smoking status, etc.

Dataset Introduction

The data for this project will be from here. It is a CSV file that contains 5110 observations with 12 attributes(11 features and 1 label).

Approach to solution

Data understanding and preparation

First, we'll go through the dataset and implement exploratory data analysis on the dataset to have a general view of the dataset. While analyzing the data, we will figure out whether the feature is useful in prediction and to what degree the predictions or, say, labels depend on each feature. We can also learn the correlation between each feature during the analysis. Moreover, the exploration will help us to determine whether we need to introduce more features in some ways, drop some redundant features, or apply l1/l2 norm regularization to avoid over-fitting. Last but not least, in this phase, we will identify data quality problems if there are any. Clean up the missing values, manage significant outliers and identical entities.

Model training

After we have the dataset cleaned and prepared for modeling, we are in the phase to explore different models and do hyperparameter selection. In this project, we will be implementing several supervised learning models, which are listed as follows:

- Logistic Regression
- SVM
- K-nearest Neighbors classifier
- Decision Tree classifier
- Random Forest classifier

- Naive Bayes classifier

For each model, we will be applying various combinations of hyperparameters to the model, so that we are able to determine the hyperparameters that perform the best on predictions, and cross validation. By comparing the results and measurements from different models, we can find the model that performs the best.

Assessment methodology

To evaluate the models and results from the models, we will plot the graphs for different measurements, such as roc curve, accuracy score, confusion matrix, recall score, precision score, and f1 score, etc.
During the cross validation phase, the various hyperparameters will be fed to the model and then check the performance of each model. We are going to use GrideSearchCV and RandomizedSearchCV to control the cross validation strategies.

Related work

1. **Analyzing the Performance of Stroke Prediction using ML Classification Algorithms**
   The authors use five accuracy criteria to assess the performance of six alternative models. In this study, the classification algorithms Logistic Regression, Support Vector Classification, Decision Tree Classification, Random Forest Classification, and Navie Bayes Classification were compared, and used the Receiver Operating Characteristic curve, Accruacy Score, Precision Score, Recall Score and F1 Score to evaluate the performance of various models. This study found that Navie Bayes performed better than other algorithms after model development.

2. **Early Stroke Prediction Using Machine Learning**
   By comparing Decision Tree Algorithm, Random Forest Algorithm, Naive Bayes Algorithm, Multi-layer Preceptron Algorithm and JRip Algorithm, authors found it is possible to forecast stroke prediction using historical data mining approaches. The highest accuracy they have achieved is 98.84% by using Random Forest algorithm.