Assignment 2

In the folder, you can find two datasets, 'train.csv' and 'test.csv'. The purpose of this assignment is to help you familiarize yourself with the classification methods we have introduced in the class. Please follow the steps below to submit the deliverables:

(1) In the 'train.csv', each observation represents one individual, who is indexed by 'index' and has a lot of features, including 'age', 'sex', 'race', 'education'. Features 'f1', 'f2', 'f3', 'f4' and 'f5' are categorical features, while 'f6', 'f7', 'f8', 'f9', 'f10' are continuous features. 'y' represents the outcome variable you are going to classify.
(2) The format of the 'test.csv' is the same as 'train.csv', except that the column 'y' is missing.
(3) Please develop classification models using 'logistic regression', 'support vector machine', 'adaboost', and "gradient boosting". Please feel free to choose the parameters needed on your own for each method.

Your deliverables should include:
(1) Your scripts for the classification. Please use jupyter notebook to do the submission. Please sort them out in an organized way for easy grading.
(2) A summary table (which should be included at the end of your jupyter notebook) to show the performance of these four methods (can use the metrics: f1-measure, precision, recall). Please also denote which method you are using to do the prediction for the test data.
(3) A csv file to show the predictions on the test data. Please include columns 'index', 'y', where 'y' stores your prediction values.

The grading will be given based on the following points:
(1) Whether you can run through the classification methods successfully, and provide a performance comparison
(2) Your prediction performance on the test data.

[1] Due Time: 11:59 PM, March 17, 2023
[2] File name: Metric Number. The name format is the same for every file in your submission. Submissions which don't follow this formatting will not be graded.