MCP261 IE Lab I: Due 11:59 PM March 19, 2024

Exercise 7: Analytics II: Introduction to Machine Learning

Please submit your files in a zipped file. The zipped file should be named as follows: " EntryNum_Name_ Ex7.zip". The zipped file should contain your Python script, named "EntryNum_Name_Ex7.py" and the data file for the machine learning problem. All submissions will be checked for evidence of plagiarism. Students whose submissions are found to have evidence of plagiarism will be subject to, at minimum, losing all marks for the exercise.

1. (10 marks) For the dataset provided, as discussed in class, perform the following operations. Use a random_state value of 20 in all applicable functions (e.g., clustering, train_test_split, SMOTE, classification algorithm).
    a. Perform K-means clustering on the tot_deaths_pm column (the label), with the number of clusters set to be 2. Use the clustering results to create labels for the dataset. As output for this step, your code should write the labelled dataset to a new excel file called "dsml.xlsx" in the same folder where the code is located. You may use the *pandas to_excel* method for this purpose.
    b. Split the new labelled dataset into train and test subdatasets with a split of 0.15.
    c. Apply the SMOTE oversampling method to the training dataset alone (use the *imblearn* package for this, and specify only the random_state attribute for the SMOTE method, let all others take their default values). As output for this step print the dimensions of the training set before and after oversampling.
    d. Scale the dataset so that all features lie in the range [0,1]. Perform scaling separately on the training and test datasets.
    e. Apply the random forest (number of estimators = 100), artificial neural network (hidden layers = (50,50,10), relu activation function, and max_iter = 500) and the logistic regression classifiers to the preprocessed dataset. Print the classification report, AUC score and sklearn's balanced_accuracy_score for each classifier.

2. (5 marks) Perform hyperparameter optimization for the random forest and neural net classifiers used in part **e** of question 1. The hyperparameter optimization must be performed via 5-fold cross-validation on the training dataset, and using the sklearn balanced_accuracy_score performance metric. The following search spaces may be used. The output of the hyperparameter optimization exercise must be an excel file called "hptoutput.xlsx". Write the hyperparameter set and the corresponding performance measure value to a different sheet for each classifier (e.g., sheet names can be 'random_forest' and 'neural_net'). Print as output the optimal hyperparameter set for each classifier based on maximum cross-validation performance.
    a. Random forest: number of estimators ranging from 50 to 500 in steps of 50.
    b. Neural network: number of layers: 2 or 3; neurons in each layer (number of neurons must be the same in each layer – e.g., valid combinations are (50,50) or (60,60,60) and not (50,60)): 50 to 80 in steps of 5; activation function: 'sigmoid' or 'relu'.

Your code should be well commented and your output should be easy to interpret.