Air University

Final-semester Examinations

Course Name: Data Science Date: Jun. 03, 2024 Instructor: Dr. Mehdi Hassan

Time: 9:30AM-11:30AM Total marks: 75 Mode: Closed Book

Submission: Online via GCR

Instructions: There are four problems given to you and it is expected to solve these problems and submit it to GCR. In case of any copying content, the AU Unfair Means Case policy will be applied.

Task-1:

1. Consider the student's performance training data (Student Performance - Training Set.csv). You are required to perform the pre-processing steps. The last column is of class labels.
2. Split the data into 80%:20% for training and testing.
3. Perform the following classifiers:
   a. Decision Tree
   b. Random Forest
   c. Bagging
   d. Boosting
   e. SVM
4. Print the confusion matrix of each algorithm (both training and testing). Find Accuracy, Sensitivity, Specificity, Precision, Recall, F-Score of all models and make a table in which these results need to be mentioned.
5. Print ROC curves of all classifiers (both training and testing) on single graph. Note: training and testing graphs should be separate.
6. Write a one-page discussion and analysis about results (minimum of 600 words).
7. Submission will be a Jupyter notebook file in which the code and its output must be visible and a CSV file of predicted values.

Task:2:

1. Consider the data of 'Adult Income Classification - Training Data.CSV'.
2. Do the required pre-processing steps.
3. Predict the income of a person using the following models.

   a. SVM (all kernels i.e. Linear RBF, and Gaussian).
   b. Find Accuracy, Sensitivity, Specificity, F-Score, and MCC of the classifiers.
   c. Plot ROC of these three kernels in a single graph.
4. Write a one-page discussion and analysis about results (minimum of 600 words).
5. Submission will be a Jupyter notebook file in which the code and its output must be visible and a CSV file of predicted values.

Task-3

Use cluster analysis to identify the groups of characteristically similar schools in the College Scorecard dataset (CollegeScorecard.CSV).

Considerations:

- **Clustering Algorithm** K-means is a powerful and recommended clustering algorithm, but the choice of clustering technique(s) is yours.
- **Data Preparation** How will you deal with missing values? Categorical variables? Feature intercorrelations? Feature normalization or scaling? Dimensionality reduction? These are the (sometimes subjective) questions you need to figure out as a data scientist. It's highly recommended to familiarize yourself with the dataset's dictionary and documentation, as well as the theory and technical characteristics of the algorithm(s) you're using.
- **Hyperparameters** How will you set the parameters -- the algorithm's knobs and dials, so to speak -- in order to achieve valid and useful output?
- Interpretation Is it possible to explain what each cluster represents? Did you retain or prepare a set of features that enables a meaningful interpretation of the clusters? Do the compositions of the clusters seem to make sense?
- **Validation** How will you measure the validity of your clustering process? Which metrics will you use and how will you apply them? Hint: Davies Bouldin Index.
- **Clusters: Do the clusters ranges from 2 to 5 and plot the DBI for every run.**
- **Important Note** This is an open-ended assignment (as many or most real-life data science projects are). Your only constraints are that you must use the data provided, execute high-quality and justifiable clustering technique, provide your rationale for the decisions you made, and ultimately produce meaningful cluster labels.

Deliverables:

- An array of cluster labels corresponding to *UNITID* (the unique college/university I.D. variable). *Note*: Due to the presence of missing data, some observations may be ommitted prior to clustering.
- The code you wrote.
- A brief discussion of the process and rationale for the technique(s) you decided to use.
- A brief explanation (interpretation) of the clusters.

Task 4:

1. Consider the data 'Forest_Fire.CSV'.
2. You are required to perform the pre-processing steps.
3. Split the data into 80%:20% (training and testing) ratio.
4. Perform multi-regression analysis on the given data. The last column 'area' is the target.
5. Compute RMSE on both training and testing.
6. See which variables you can omit based on the 'P' values.
7. Plot the Residual plots of both training and testing data.
8. Do step 7 on both before and after omission. You need to apply t-test on the predictions of before and after omission with null hypothesis that there is no difference in predictions.
9. Write a one-page discussion and analysis about results (minimum of 600 words).
10. Submission will be a Jupyter notebook file in which the code and its output must be visible and a CSV file of predicted values.