CS 457 - Homework Assignment 9: Classification

Due Date: Monday, November 06 at 11:59 pm

Purpose:
Demonstrate understating of Classification technique for prediction and recommendation

Points: 100

Deliverables: Submit ipynb code file along with your answer

Use the dataset `CreditCardData.csv` and `mcdonalds.csv`

- Perform analysis on the following questions**.** Make sure to include interpretation of each result including accuracy, confusion matrix and other necessary information to support your answer

Classification using `CreditCardData.csv`

1. Need to clean the data. Remove all the rows which have any missing values in any column (missing data is represented as ? and not nan/null so you need to keep this in mind)
2. Replace the **Approved** column (response variable) values from -/+ to 0/1 or No/Yes based on your preference.
3. Create a train and test set after cleaning the data. Use 30% (0.3) records for test set. Use the same train and test set for all your analysis with different classifiers.
4. In your code, set the seed after you read the data. This will keep your data and calculation consistent throughout the analysis irrespective of multiple runs. See the example code for the class `random_state=99`
5. Create a classification model to predict Approved status **using Decision Tree**. Visualize the decision tree. Interpret the decision tree. Discuss which attributes are important and which are not important.
6. Perform the Tree Pruning Analysis and evaluate the results. Visualize the tree after pruning. Discuss the tree and overall results before and after pruning.
7. Create a classification model to predict Approved status **using RandomForest**. Include all the analysis steps including variable importance plot. Try atleast 5 different values of n_estimators (number of trees) and compare the classification error. Pick the best model (based on n_estimators) for your final comparison.
8. Report the comparison between Decision Tree and RandomForest in terms of classification performance.

Recommendation using `mcdonalds.csv`

1. Build a decision tree to recommend similar food based on their attributes. Interpret your decision tree and tell some interesting insights from it. You do not (necessarily) need to pick all the food (rows) and attributes (columns). Just pick 3 similar food items of your choice and explain/interpret your recommendations.
   Note: You need to remove restaurant column first from the data before creating a decision tree.