**PCC 5: Supervised Learning SVM / DT / RF**

Due April 6th in Canvas, 2020 5:00 pm

For this PCC, you are asked to demonstrate your knowledge of supervised learning techniques by analyzing one of the two datasets below and another dataset of your choosing (you are asked to provide the dataset and the URL reference). On both datasets, you should use Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF).

This is an individual assignment.

Your deliverable is a PPT that includes a description of your algorithms, your results, and sensitivity analyses.

**Dataset 1a: Train Transportation**

Spain Local Train Transportation: Spain Rail Transportation.csv

1. Create a statistical summary of the data set
2. Create a plot to show what is the average time between ORIGIN and DESTINATION
3. Group all the data per month, per week. Make a plot to display:
   1. The Number of tickets bought
   2. The Trend of the average travel time
   3. The Trend of the price
   4. A bar plot to see the distribution of destination, train type, train class, and fare.
4. Evaluate and compare SVM, DT, and RF, to determine what fare will paid, based on the feature you consider relevant. Which one of these three methods gives the best result? What are the best parameters for each algorithm?
5. Perform a sensitivity analysis on the effects of training % vs. testing %
6. Make a summary results table showing the accuracy score, confusion matrix, and classification report for the methods compared.
7. Randomly create 10 instances based on the features used in the model, and then use the model to predict the quality for bread.

**Dataset 1b: Bread Quality**

Composition of Bread: Bread quality.csv

1. Create a statistical summary of the data set.
2. Create an appropriate histogram for each of the features you consider relevant (at least 3).
3. Analyze the data using SVM, DT, and RF, to predict what expected QUALITY for bread would be obtained. Which one these three methods give the best result? What are the best parameters for each algorithm?
4. Perform a sensitivity analysis on the effects of training % vs. testing % for the best parameters obtained per each method.
5. Create a boxplot comparing the error from the three ML Algorithms you tested.
6. Make a summary results table showing the accuracy score, confusion matrix, and classification report for the methods compared.
7. Randomly create 10 instances based on the features used in the model, and then use the model to predict the quality for bread.

**Your Dataset**

1. Create a statistical summary of the data set.
2. Create an appropriate histogram for each of the features you consider relevant (at least 3).
3. Analyze the data using SVM, DT, and RF. Which one these three methods give the best result? What are the best parameters for each algorithm?
4. Perform a sensitivity analysis on the effects of training % vs. testing % for the best parameters obtained per each method.
5. Create a boxplot comparing the error from the three ML Algorithms you tested.
6. Make a summary results table showing the accuracy score, confusion matrix, and classification report for the methods compared.
7. Randomly create 10 instances based on the features used in the model, and then use the model to predict the quality for bread.