

**ISM 6136 – Datamining/Predictive Analytics**

**Class Assignment 5**

**5 points**

**TASK: Performing predictive analytics using Classification and Regression Trees in RapidMiner and XLMiner**

**Part 1**

**Perform the following data mining steps using classification tree algorithm using RapidMiner and predict which of the wineries A, B or C does the wine in the new data belong to.**

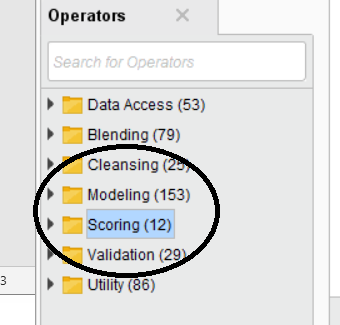
Follow the datamining steps below:

1. Understand the problem and purpose of data mining task
2. Import the dataset into RapidMiner
3. Explore, clean and preprocess data

Cleanup or do not select any column that is not a predictor

* + 1. Check ‘Replace errors by missing values’
    2. Check Statistics and look for any missing values (if yes then you will have to add a replace with the ‘Replace Missing Values’ operator

1. Design your process using appropriate operators. Provide screen shot of the overall design.



1. **Build two models** – save each of the models into a separate ‘process’
2. **Select the best model - Provide screen shots of your model selection criteria**
3. Apply New Data Scoring to the best model process
4. Provide screen shot of the prediction results on the New Data
5. **Explain and provide screen shots of steps f) through i). Please remember to provide screen shots for the ‘process’ showing all operators and their performance matrix (confusion matrix). Also provide screen shot of the ‘Decision Tree’ created by the ‘best’ model and explain which settings were changed to get this best model. Any other settings that were changed to get other models.**

**Upload only the Word document with explanation and screen shots of the results.**

**Part 2**

**Perform the following data mining steps using classification tree algorithm using XLMiner and predict which of the wineries A, B or C does the wine in the new data belong to.**

Follow the datamining steps below:

1. Understand the problem and purpose of data mining task
2. Obtain the dataset for analysis – Wine.xls
3. Explore, clean and preprocess data
   1. Cleanup any column that is not a predictor
   2. Perform ‘Missing Data Handling’ *(Delete any record that is corrupt)*
4. Reduction of data dimension (if needed)
5. Partition data
6. Choose the data mining techniques/algorithms – Classify > Classification Tree and **build 2 models**
7. Follow all the steps of creating a model as shown in Lecture 7 slides (16,17, 18)
8. Interpret the results and depending on the model selection criteria choose the **best model**
9. For the best model - Provide screen shots and explain the comparison of the Full grown, Best Pruned and Minimum error tree in term of the predictors selected in them**. Can the best prune tree be used for dimension reduction? How many dimensions did it reduce?**
10. Deploy **best model** on the new data and explain your prediction results
11. Submit the **Excel workbook and Word document with explanation and screen shots for steps h) through j).**