

**ISM 6136 – Datamining/Predictive Analytics**

**Class Assignment 7**

**5 points**

**TASK: Performing predictive analytics using Neural Nets/ANN in XLMiner OR RapidMiner**

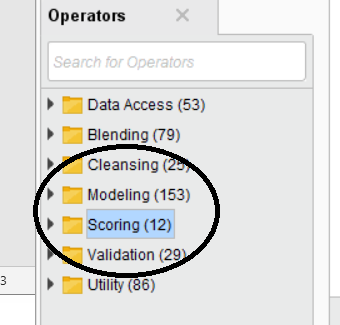
**Perform the following data mining steps (CLASSIFICATION) in XLMiner**

1. Follow the datamining steps below:
2. Understand the problem and purpose of data mining task
3. Obtain the dataset for analysis – **Breast Cancer Diagnosis.xls**
4. Explore, clean and preprocess data
   1. Cleanup any column that is not a predictor
   2. Perform ‘Missing Data Handling’
   3. Any categorical variables conversion needed – check and remember to perform during modeling
5. Reduction of data dimension (if needed to get another model)
6. Partition data
7. Choose the data mining techniques/algorithms – **Classify > Neural Networks** > **Automatic or Manual** and **build 4 models - (Remember to Rescale** (use standardization) **the dataset)**
8. Follow the steps of creating a model as shown in Lecture 9 slides
9. Interpret the results and depending on the model selection criteria choose the **best model**
10. Deploy **best model** on the new data and explain your prediction results (how many records/instances are Malignant/Benign. Paste a screenshot of your prediction for new data.
11. Submit the **Excel workbook and this word document with explanation/screenshots for steps c) through i).**

**OR**

**2. Follow the datamining steps below using RapidMiner**

1. Understand the problem and purpose of data mining task
2. Import the dataset into RapidMiner
3. Explore, clean and preprocess data
4. Cleanup - 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
   3. Don’t need to normalize as RapidMiner takes care of it *(there is a normalize check by default on the Neural Net operator settings).* If you add ‘normalize’ operator it will mess up the results.
5. Design your process using appropriate operators. Provide screen shot of the overall design.



1. **Build four models** – save each of the models into a separate ‘process’

Provide screen shots of each of the model settings– to show me difference in each model – You can even present this in Table form.

1. Select the best model - Provide screen shots of your model selection criteria – confusion matrix (accuracy, precision, recall) and AUC, ROC chart.
2. Provide screen shot of the ‘ANN Diagram’ - you get it by connecting ‘mod’ to ‘res’ on the Cross Validation operator.
3. Apply New Data Scoring to the best model process – provide screen shot of the Design process
4. Provide screen shot of the prediction results on the New Data
5. **Explain and provide screen shots of steps f) through j).**