

**ISM 6136 – Datamining/Predictive Analytics**

**Jacob Perrone**

**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

Graphical user interface, text, application, Word

Description automatically generated

For this dataset, all the features were numerical, and only 2 records had missing values. The instances with missing values were deleted since there was so few of them.

1. Reduction of data dimension (if needed to get another model)

Not done for this assignment, but could have done PCA to find the best features and only use those for the NN instead of using all 30 features.

1. Partition data

Table

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

I chose two partitioning levels for the NN’s, a 60/40 split and a 80/20 split. For other models (in other assignments) an 80/20 split normally produced better results than 60/40 splits and I was curious if the same would be for NN.

1. Choose the data mining techniques/algorithms – **Classify > Neural Networks** > **Automatic or Manual** and **build 4 models - (Remember to Rescale** (use standardization) **the dataset)**

**For getting the four models, I chose to do an automatic NN which took longer to build but created vastly more NN options of layers and nodes in each layer and there results. By doing this, I sorted those models by lowest error % to choose the best models. I chose 2 from each partitioning (shown in table).**

1. Follow the steps of creating a model as shown in Lecture 9 slides

After obtaining NNC\_Output and NNC\_Output3, which are the results of running the automatic version of NN in the classify section. After sorting by lowest error %, I chose the two best models from each (60/40 and 80/20 partitioning for the respective output tabs). By clicking on the net name, it brough up the selection criteria similar to how a manual NN would be made, I made sure that the data was rescaled using standardization (also done in previous step), also made sure neural net weights was checked and that the detailed reports had the lift charts selected as well. That was done for each model chosen on each output tab.

1. Interpret the results and depending on the model selection criteria choose the **best model**

Table

Description automatically generated

Graphical user interface, chart, line chart

Description automatically generated

**First, for selecting the best model I looked at the accuracy of training and validation, which model 1 had the highest of. The validation error % was also the lowest as a result of the high accuracy, and model 1 also had a high F1 score. After that I looked at the lift chart and the AUC for model 1 validation (shown above). Model 1 had a very high AUC of 0.99444 which is very good. The lift chart also shows a good lift from the actual line (red).**

1. 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.

Table

Description automatically generated

1. Submit the **Excel workbook and this word document with explanation/screenshots for steps c) through i).**