**Assignment 7 Answers –Highlighted in yellow**

**Used XLMiner**

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

**1.** Follow the datamining steps below:

**a)** Understand the problem and purpose of data mining task

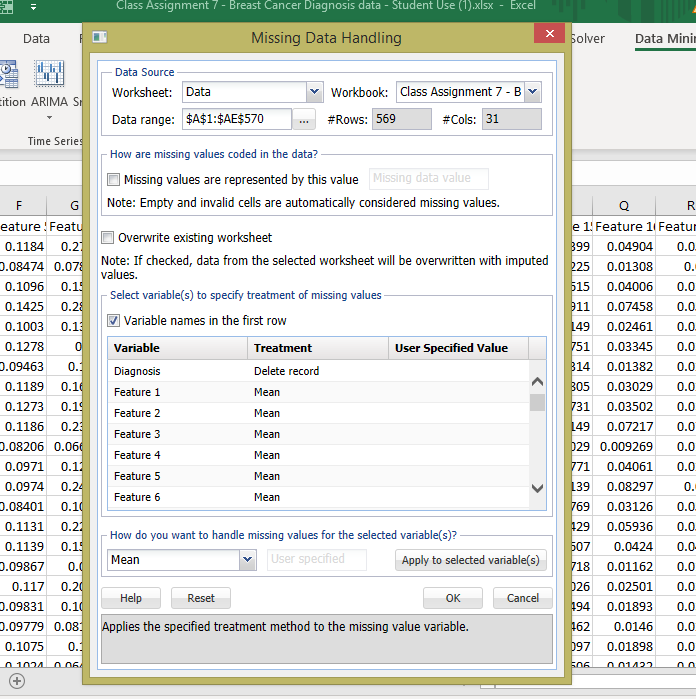
**b)** Obtain the dataset for analysis – **Breast Cancer Diagnosis.xls**

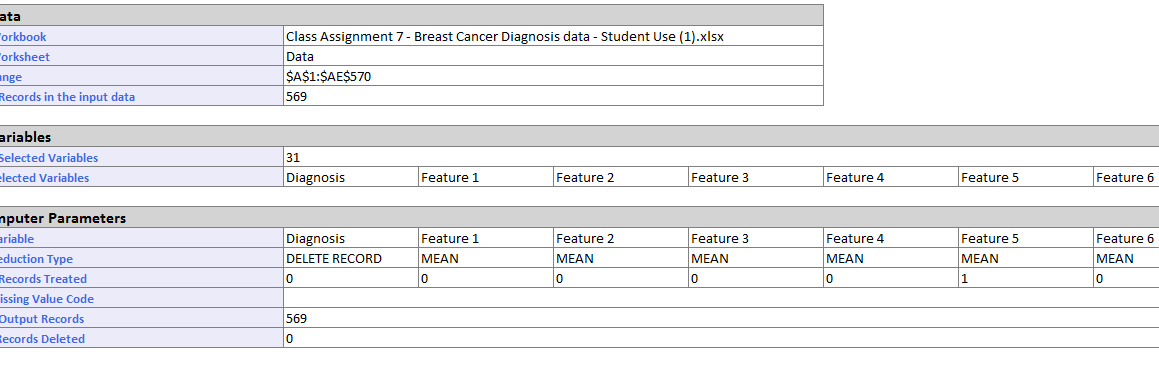
**c)** Explore, clean and preprocess data

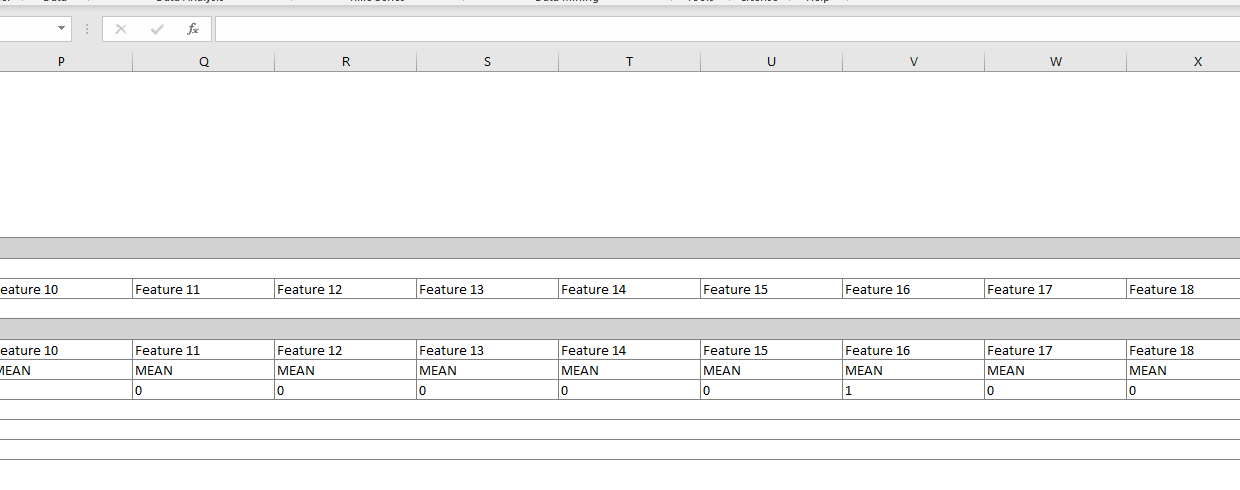
i. Cleanup any column that is not a predictor – Column A (No of instances) and Column B(ID) are not predictors, hence deleted.

ii. Perform ‘Missing Data Handling’

Delete for outcome/Diagnosis and replace with “mean” for features. Two features (feature 5 and feature 16) data was replaced per snapshot below.







iii. Any categorical variables conversion needed – check and remember to

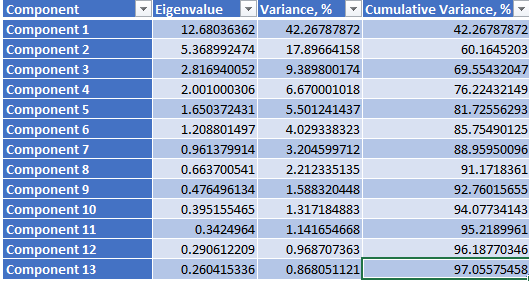
perform during modeling

No categorical input data to be converted.

**d)** Reduction of data dimension (if needed to get another model)

Applied PCA for dimension reduction in Model 2, Model 3 and Model 4

This included 97% of the data.



**e)** Partition data

Model 1,2 and 4: 60/40

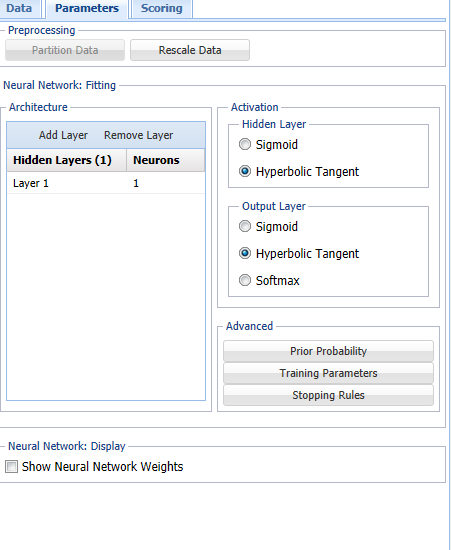
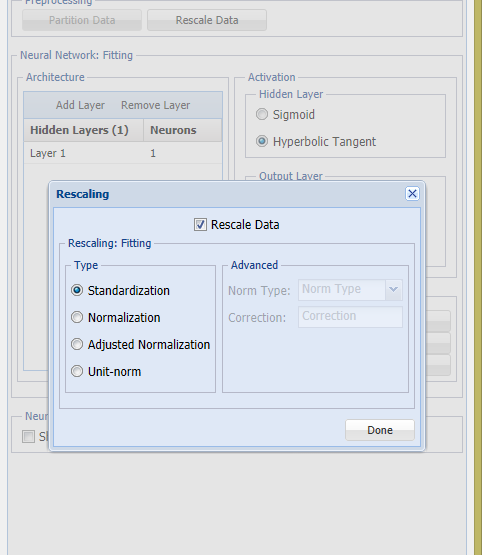
Model 3: 80/20 -🡪 led to decreased accuracy

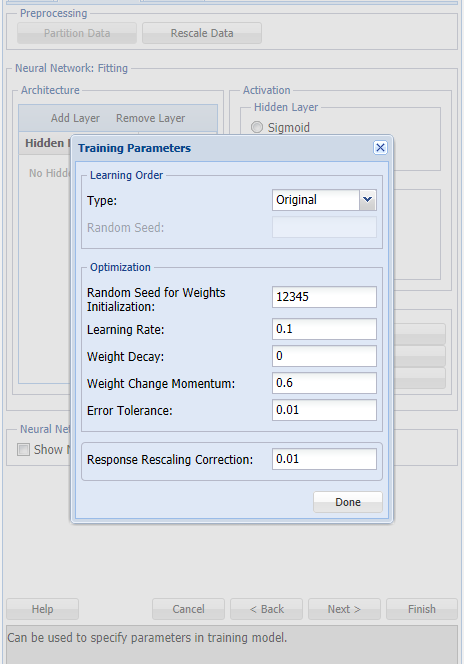
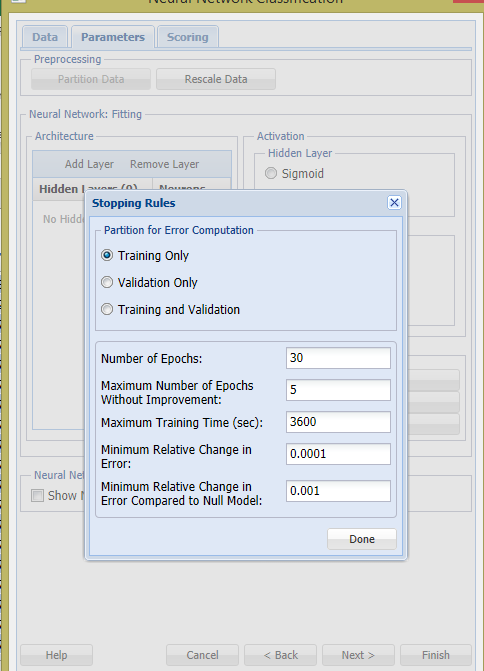
**f)** Choose the data mining techniques/algorithms – **Classify > Neural Networks** >

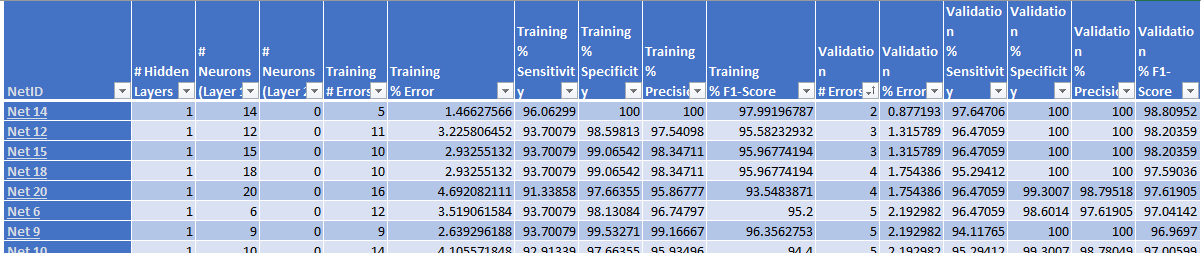
**Automatic or Manual** and **build 4 models - (Remember to Rescale** (use

standardization) **the dataset)**

**Criteria changed in the following places (these were settings in Model 1)**





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criteria** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| **Features** | Manual  30 dimensions  Partition: 60/40  Layer: 1, Neuron: 1  Epoch: 30 | Manual  **13 dimensions**  Partition: 60/40  Layer: 1, Neuron: 1  Epoch: 30 | Manual  13 dimensions  **Partition: 80/20**  Layer: 1, Neuron: 1  Epoch: 30 | **Automatic**    30 dimensions  Partition: 60/40  **Layer: 1, Neuron: 14**  Epoch: 30 |
| **% Error** | Training: 5.57%  Validation: 3.95% | Training: 26.98%  Validation: 27.19% | Training: 25.05%  Validation: 35.08% | Training: 1.47%  Validation: 0.88% |
| **Accuracy** | Training: 94.43%  Validation: 96.05% | Training: 73.02%  Validation: 72.81% | Training: 74.95%  Validation: 64.91% | Training: 98.53%  Validation: 99.12% |
| **Sensitivity** | Training: 0.93  Validation: 0.94 | Training: 0.61  Validation: 0.61 | Training: 0.63  Validation: 0.47 | Training: 0.96  Validation: 0.98 |
| **Precision** | Training: 0.92  Validation: 0.95 | Training: 0.64  Validation: 0.64 | Training: 0.67  Validation: 0.57 | Training: 1  Validation: 1 |
| **F1** | Training: 0.93  Validation: 0.95 | Training: 0.63  Validation: 0.63 | Training: 0.65  Validation: 0.51 | Training: 0.98  Validation: 0.99 |
| **Lift Chart (validation)** |  |  |  |  |
| **ROC (validation)** |  |  |  |  |
| **AUC (validation)** | **0.997** | **0.796** | **0.711** | **0.998** |

**g)** Follow the steps of creating a model as shown in Lecture 9 slides

**h)** Interpret the results and depending on the model selection criteria choose the **best**

**model**

**The features adjusted for each model are displayed in the table above.**

As can be seen from the above table, Model 4 (Neural Network- Automatic) performed the best because the validation error was low and accuracy/sensitivity/precision and F1 were all high. The training and validation errors were close to each other. The lift chart showed the greatest area between the lift curve and the base line for model 4. The ROC curve was closest to the top left corner of the graph for model 4.

Hence, it was concluded that model 4 performed the best.

**i)** Deploy **best model** on the new data and explain your prediction results (how many

records/instances are Malignant/Benign

Scoring run on model 4 shows the following predictions – four malignant cases and 16 benign cases.

|  |  |  |  |
| --- | --- | --- | --- |
| **Record ID** | **Prediction: Diagnosis** | **PostProb: B** | **PostProb: M** |
| **Record 1** | M | 0 | 1 |
| **Record 2** | B | 0.992508232 | 0.007491768 |
| **Record 3** | B | 0.771282285 | 0.228717715 |
| **Record 4** | B | 0.785502792 | 0.214497208 |
| **Record 5** | B | 0.818621104 | 0.181378896 |
| **Record 6** | B | 0.554692438 | 0.445307562 |
| **Record 7** | B | 0.909608779 | 0.090391221 |
| **Record 8** | B | 1 | 0 |
| **Record 9** | B | 0.898947981 | 0.101052019 |
| **Record 10** | B | 0.793774721 | 0.206225279 |
| **Record 11** | B | 0.782049651 | 0.217950349 |
| **Record 12** | B | 0.88593323 | 0.11406677 |
| **Record 13** | M | 0 | 1 |
| **Record 14** | B | 0.682063848 | 0.317936152 |
| **Record 15** | M | 0 | 1 |
| **Record 16** | M | 0.398574596 | 0.601425404 |
| **Record 17** | B | 0.72014878 | 0.27985122 |
| **Record 18** | B | 0.805628579 | 0.194371421 |

**j)** Submit the **Excel workbook and this word document with explanation/screenshots**

**for steps c) through i).**

**The best model and the scoring tabs are highlighted in yellow.**