

**ISM 6136 – Datamining/Predictive Analytics**

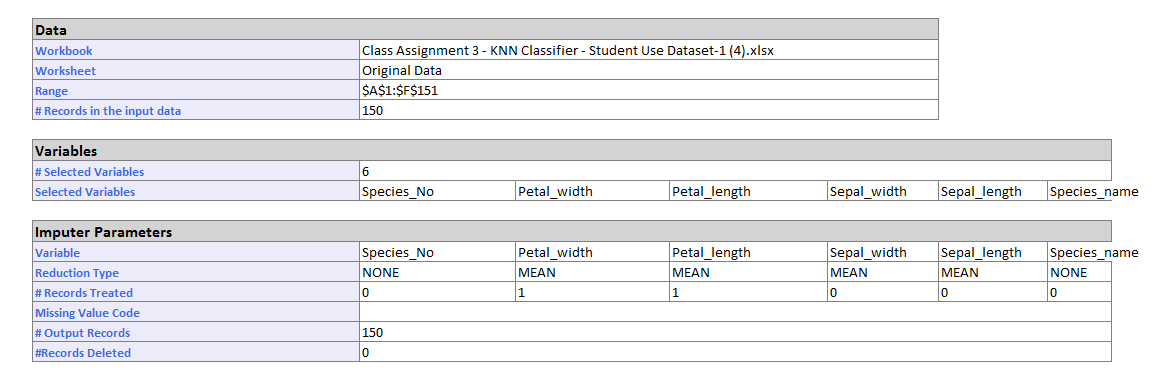
**Class Assignment 3**

**5 points**

**TASK: Performing classification using the k-NN classifier.**

1. **Question: Provide explanation of steps c) through h) on a word document.** Regarding the model selection explain the ‘full’ model selection criteria (Slide# 16 of Lecture 5) you followed and how you compared each of the models. You can present it in tabular form too with results from each model.

c) **Missing Data Handling** – Cleaned data using Transforming (Handling Missing Data).Two substitutions were made with mean as shown in the following snapshot.



1. **Determine the appropriate data mining task**

Appropriate data mining task used “classification – k nearest neighbors”.

1. **Partition data (supervised tasks)**

The data was partitioned as shown in the table below.

1. **Choose the data mining techniques/algorithms**

classification – k nearest neighbors. In order to improve accuracy, k was increased to 9 (maximum odd value allowed by the software).

1. **Build the model by interpreting results of algorithms**

The following 5 models and the results are shown below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model #** | **Partition** | **k** | **Accuracy Rate** | **Error Rate** | **Diagonal** | **Non-Diagonal** |
| # 1 | 60/40 | 5 | Training – 95.6%  Validation – 96.7% | Training – 4.4%  Validation – 3.3% | Training – 86  Validation – 58 | Training – 4  Validation – 2 |
| # 2 | 75/25 | 5 | Training – 96.5%  Validation – 100% | Training – 3.5%  Validation – 0% | Training – 109  Validation – 37 | Training – 4  Validation – 0 |
| # 3 | 75/25 | 7 | Training – 94.7%  Validation – 100% | Training – 5.3%  Validation – 0% | Training – 107  Validation – 37 | Training – 6  Validation – 0 |
| # 4 | 60/40 | 7 | Training – 94.4%  Validation – 96.7% | Training – 5.6%  Validation – 3.3% | Training – 85  Validation – 58 | Training – 5  Validation – 2 |
| # 5 | 60/40 | 9 | Training – 96.7%  Validation – 98.3% | Training – 3.3%  Validation – 1.7% | Training – 87  Validation – 59 | Training – 3  Validation – 1 |

1. **Compare the models and select the best one**

The best model selected was Model #3.

**Criteria for Model Selection**

1. Training Vs Validation accuracy rate for each model – Each individual model was checked for no vast differences between training accuracy rate and validation accuracy rate.
2. Training Vs Validation error rate for each model– Each individual model was checked for no vast differences between training accuracy rate and validation error rate.
3. Validation Vs Validation Accuracy rate for model comparison- The highest accuracy rate amongst all the models was chosen – Model #3
4. Validation Vs Validation Error/misclassification rate for Model comparison – The lowest error rate amongst all the models was chosen – Model #3
5. Diagonal value comparison for model comparison- Adding the diagonals and non-diagonals of the Comparison matrix will give an idea of the best model with the highest precision for this 3 by 3 matrix.

We are looking for a model with “0” or close to 0 non-diagonal value – highest precision. As seen from the above table, Model #3 gives the lowest non-diagonal value (0) and diagonal value (37).

1. K value comparison -Model #3 has a reasonably high k value of 7 for such a small data set.
2. **Deploy ‘only’ the best model (score on new data)**

The scoring was run for model #3 results of which are shown below.

