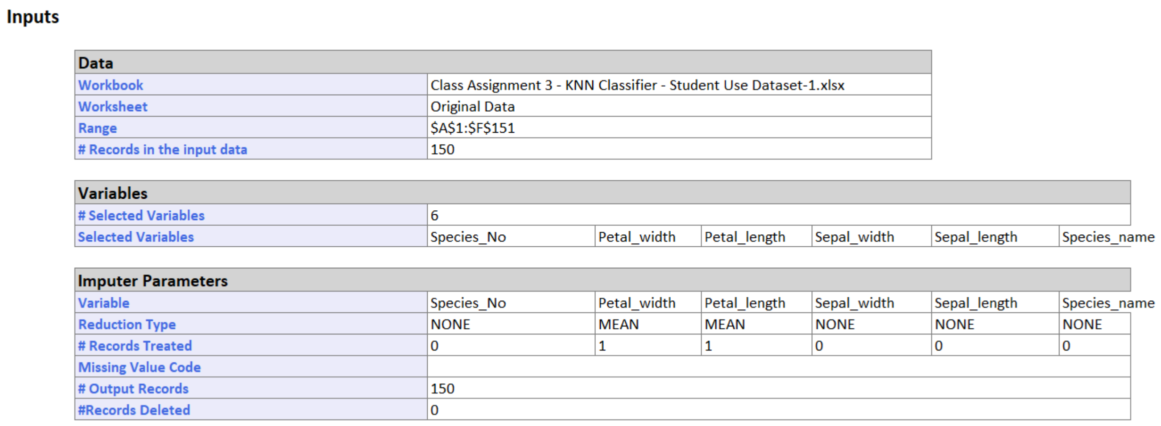
Jacob Perrone

Graded Assignment 3 Report



Inputs

Data	
Workbook	Class Assignment 3 - KNN Classifier - Student Use Dataset-1.xlsx
Worksheet	Original Data
Range	$A$1:$F$151
# Records in the input data	150

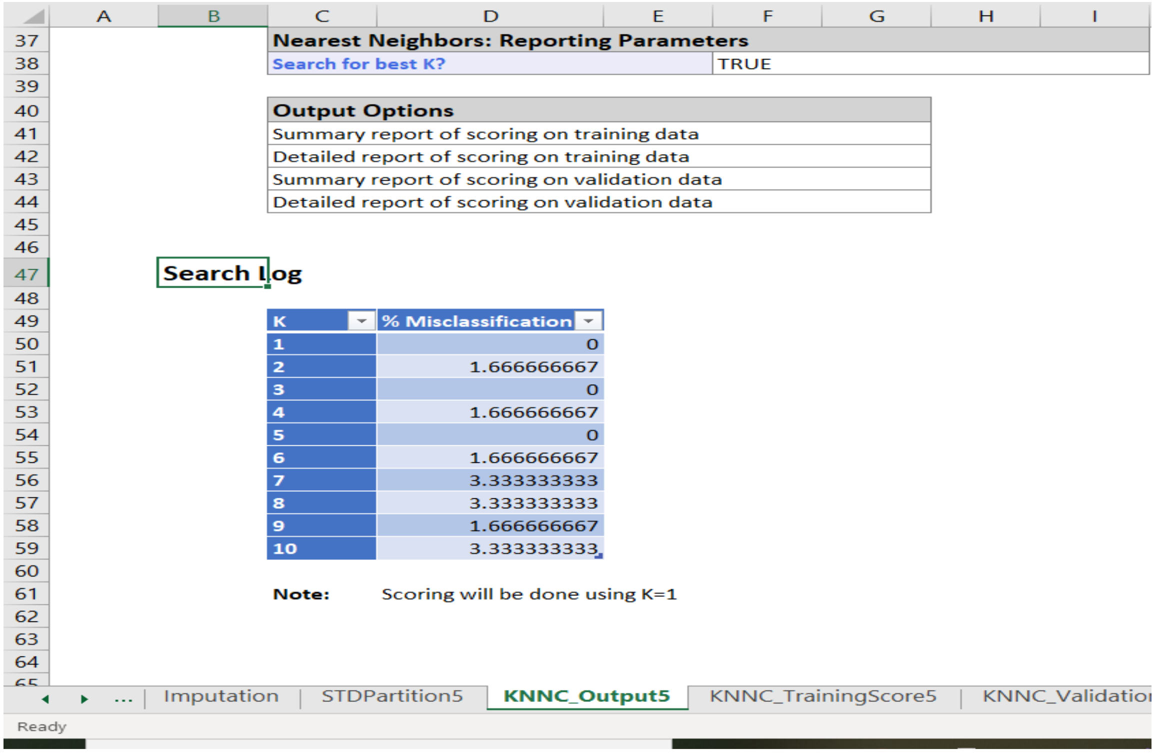
Variables						
# Selected Variables	6					
Selected Variables	Species_No	Petal_width	Petal_length	Sepal_width	Sepal_length	Species_name

Imputer Parameters						
Variable	Species_No	Petal_width	Petal_length	Sepal_width	Sepal_length	Species_name
Reduction Type	NONE	MEAN	MEAN	NONE	NONE	NONE
# Records Treated	0	1	1	0	0	0
Missing Value Code						
# Output Records	150					
#Records Deleted	0					

Part C) There were two records that had missing values. One in the petal length and the other in petal width, using the missing data handling feature and replaced the missing values using the mean to replace the missing values. No dummy values are needed as there are no categorical predictor entries.

Part D) For this data mining task, KNN will be used to predict what class of Iris based on the 4 predictors. Different partitioning will be used and compared along with different K values for nearest neighbors.

Part E-H)



Nearest Neighbors: Reporting Parameters	
Search for best K?	TRUE

Output Options
Summary report of scoring on training data
Detailed report of scoring on training data
Summary report of scoring on validation data
Detailed report of scoring on validation data

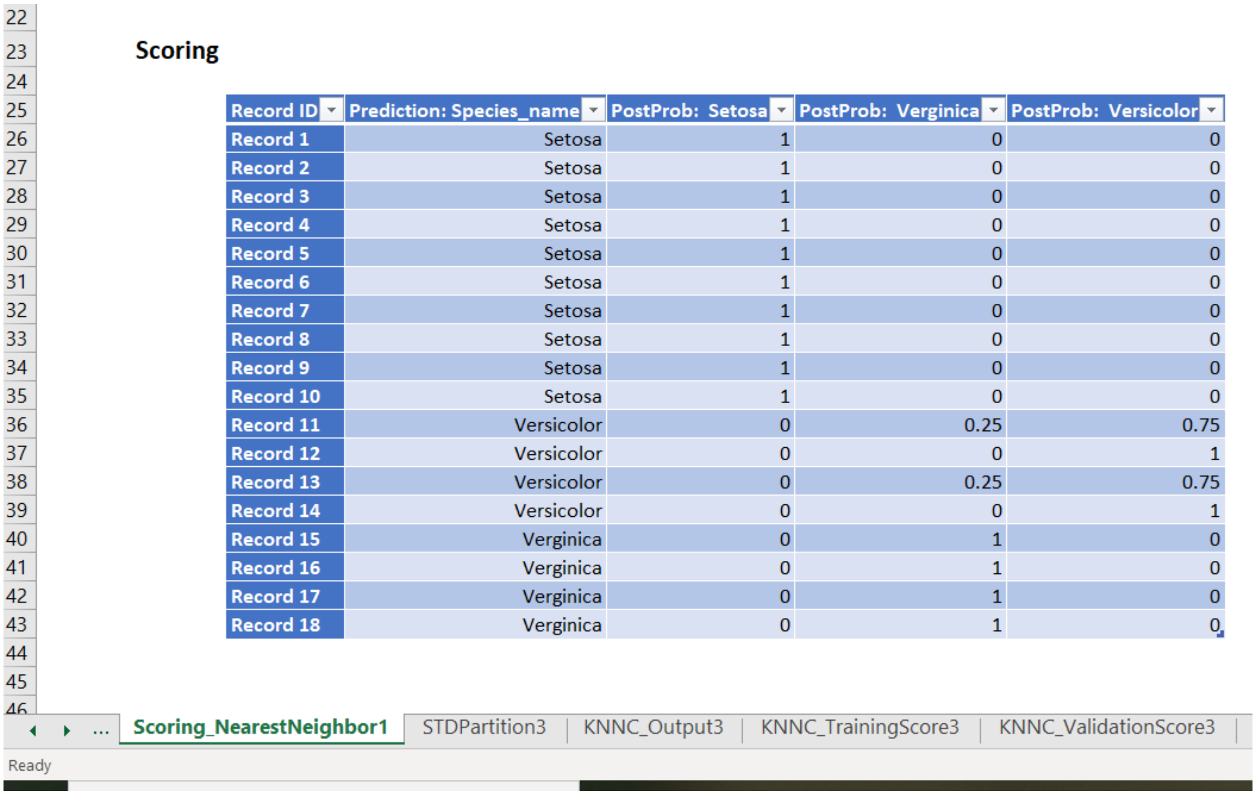
Search Log

K	% Misclassification
1	0
2	1.666666667
3	0
4	1.666666667
5	0
6	1.666666667
7	3.333333333
8	3.333333333
9	1.666666667
10	3.333333333

Note: Scoring will be done using K=1

With a 60/40, and 80/20 partition was used along with varying k values which are tabulated in table of models tab. I then ran another partitioning with 60/40 split with a search of nearest neighbors between 1 and k (maxed to 10 in XLMiner). Above is the screenshot of the results of that model for the different k values and the misclassification percentages. According to that table a 60/40 split would be the best model to use (given that is the lowest k value greater than 3 which was stated as requirement in instructions). The k values that have a zero misclassification may suffer from overfitting and a separate score tab will be run to determine .

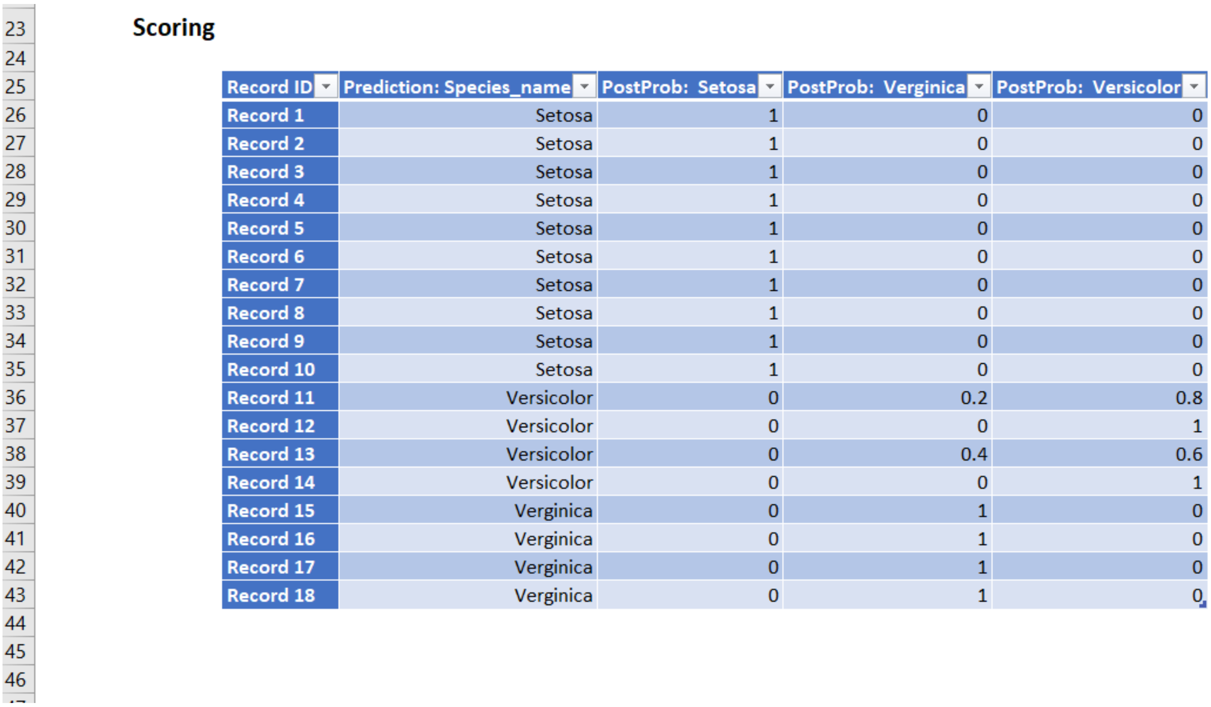
Scoring results k = 4 which is output 4 tab



Scoring

Record ID	Prediction: Species_name	PostProb: Setosa	PostProb: Verginica	PostProb: Versicolor
Record 1	Setosa	1	0	0
Record 2	Setosa	1	0	0
Record 3	Setosa	1	0	0
Record 4	Setosa	1	0	0
Record 5	Setosa	1	0	0
Record 6	Setosa	1	0	0
Record 7	Setosa	1	0	0
Record 8	Setosa	1	0	0
Record 9	Setosa	1	0	0
Record 10	Setosa	1	0	0
Record 11	Versicolor	0	0.25	0.75
Record 12	Versicolor	0	0	1
Record 13	Versicolor	0	0.25	0.75
Record 14	Versicolor	0	0	1
Record 15	Verginica	0	1	0
Record 16	Verginica	0	1	0
Record 17	Verginica	0	1	0
Record 18	Verginica	0	1	0

Scoring results k = 5 which is output 2 tab



Scoring

Record ID	Prediction: Species_name	PostProb: Setosa	PostProb: Verginica	PostProb: Versicolor
Record 1	Setosa	1	0	0
Record 2	Setosa	1	0	0
Record 3	Setosa	1	0	0
Record 4	Setosa	1	0	0
Record 5	Setosa	1	0	0
Record 6	Setosa	1	0	0
Record 7	Setosa	1	0	0
Record 8	Setosa	1	0	0
Record 9	Setosa	1	0	0
Record 10	Setosa	1	0	0
Record 11	Versicolor	0	0.2	0.8
Record 12	Versicolor	0	0	1
Record 13	Versicolor	0	0.4	0.6
Record 14	Versicolor	0	0	1
Record 15	Verginica	0	1	0
Record 16	Verginica	0	1	0
Record 17	Verginica	0	1	0
Record 18	Verginica	0	1	0

Looking at the highlighted models, I would choose k = 4 as the best model as some of the other models had perfect accuracy which led me to believe that the model might be overfitting on the training data but could also be that the model saw all examples in the training data and was able to correctly classify those in the validation set. Looking at the second scoring results (which is from the lowest k value result), the results are similar to the k = 4 model. Looking at the other models, the ones that had perfect classification I thought that those models were overfitting at first, but the dataset is small so it might be possible that the model is not overfitting because all unique training examples were seen so validation classified correctly.