Jacob Perrone

Graded Assignment 3 Report

Graphical user interface, text, application, email

Description automatically generated

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)

Graphical user interface, table

Description automatically generated

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

Table

Description automatically generated

Scoring results k = 5 which is output 2 tab

Table

Description automatically generated

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.