Decision Tree Assignment

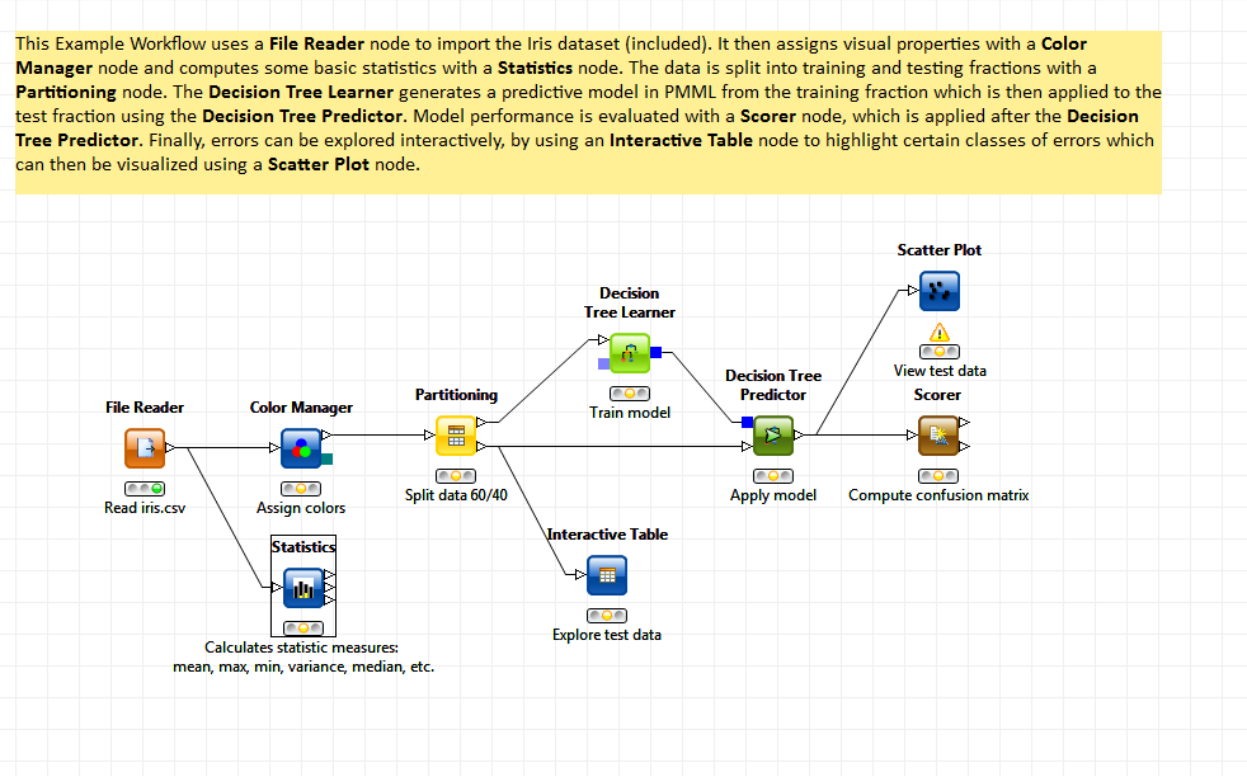
Create a KNIME workflow that utilizes the Decision Tree learning method to train a model on the Iris training data set. This data set can be found at the UCI machine learning repository.

http://archive.ics.uci.edu/ml/datasets/Iris

The Decision Tree method is available in the Node repository panel under "Mining"-> "Decision Tree" category. There are two different nodes: "Decision Tree Learner" and "Decision Tree Predictor". Decision Tree learner node creates a Decision Tree model from the input training data. The Decision Tree Predictor node applies an existing Decision Tree modes to the input data table. Additionally, the scorer node should be added at the end of the workflow in order to measure classifiers' performance.

https://www.knime.org/introduction/examples

An example of the workflow should look like the workflow below:



https://www.knime.org/introduction/examples

Task #1.

By utilizing the Decision Tree learner, predictor and scorer nodes – how well was the Model able to learn to predict Iris flowers. Hint: take a look at the confusion matrix within the scorer node. How many Iris-setosas has the Decision Tree misclassified?

Task #2.

Looking at the accuracy statistics of the scoring node on the Decision Tree model trained on the Iris data set– which class had the highest Recall and Precision?

Task #3

Create a KNIME workflow that utilizes the Decision Tree learning method to train a model on the Adult training data set. This data set can be found at the UCI machine learning repository.

http://archive.ics.uci.edu/ml/datasets/Adult

This data set was developed by Barry Becker by extracting from the 1994 Census database.
Prediction task is to determine whether a person makes over 50K a year.

Use the file reader node to read in the training data file from:

http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

Looking at the accuracy statistics of the scoring node – which class had the highest Recall and Precision?

Task #4

Change the Decision Tree learner node parameters (for the Adult training data set) from the default Gini index to Gain Ratio. By changing this parameter – did the resulting tree able to obtain a higher or lower error than the default Gini index?

Task #5

Change the Decision Tree learner node Pruning Method Parameter to MDL pruning. Is the resulting tree able to obtain a higher or lower overall prediction error as compared to no pruning alternative?