Student Number: [redacted] Name: Bryan Hoang

1. (2 marks) Use the Naive Bayes classifier as a predictor. Report results such as classification accuracy, confusion matrix, and perhaps some of the other measures we discussed.

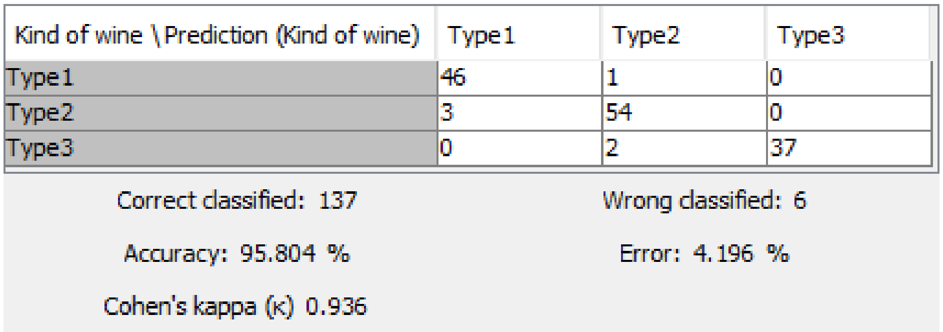
   Explore the options for binning (that is converting a numerical attribute into a categorical one - really an ordinal one - by putting the values into ranges and calling each range a bin).

   Try this with at least one attribute in KNIME, with several different choices of bins, and see what effect this has on prediction performance.

   Answer:

   For the exercise, I used the KNIME workflow, seen in Figure 6, to predict the wine's type using all the attributes, and then assessed the use of binning to improve prediction results using on Flavanoids.

   Using the Naive Bayes classifier as a predictor with a 66%/33% partitioning for training and test data led to the model achieving a prediction accuracy of ~96%. As can be observed in Figure 1, the confusion matrix shows that the model misclassified relatively few records.



Kind of wine \ Prediction (Kind of wine)	Type1	Type2	Type3
Type1	46	1	0
Type2	3	54	0
Type3	0	2	37

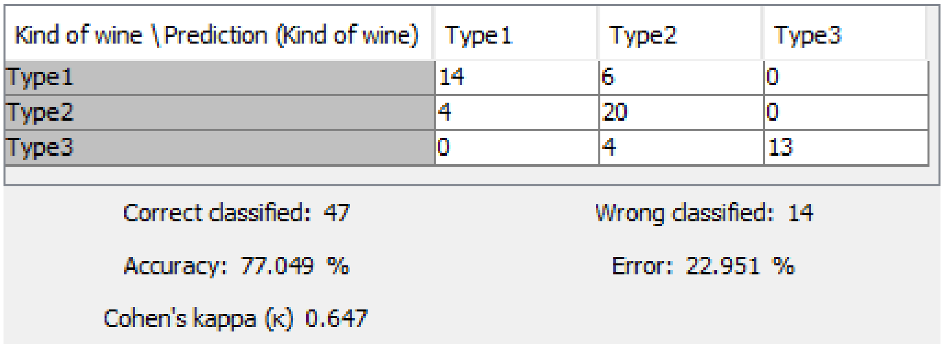
Correct classified: 137 Wrong classified: 6

Accuracy: 95.804 % Error: 4.196 %

Cohen's kappa (κ) 0.936

Figure 1: Confusion matrix of the Naive Bayes predictor using all attributes

   After filtering the data to only include the target attribute (wine type) and the Flavanoids, the model achieved a ~77% prediction accuracy, as shown in Fiture 2.



Kind of wine \ Prediction (Kind of wine)	Type1	Type2	Type3
Type1	14	6	0
Type2	4	20	0
Type3	0	4	13

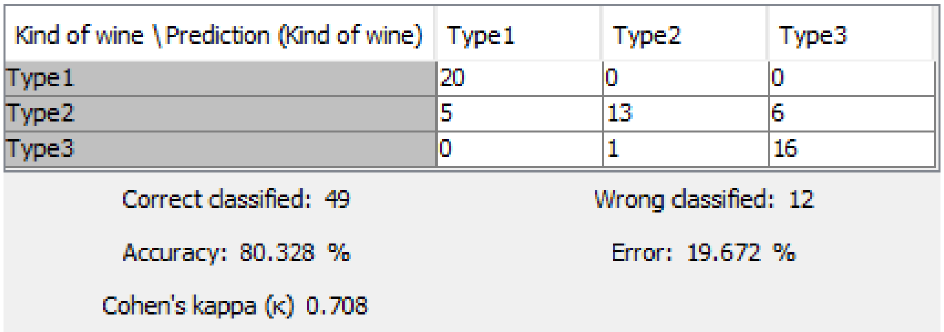
Correct classified: 47 Wrong classified: 14

Accuracy: 77.049 % Error: 22.951 %

Cohen's kappa (κ) 0.647

Figure 2: Confusion matrix of the Naive Bayes predictor using only Flavanoids

I first tried manually binning the data after viewing the sorted the csv data along the Flavanoids column using the following intervals: $[0, 1.5], [1.5, 2.3], [2.3, \infty]$. I determined the endpoints by looking for notable transitions in the wine type in the sorted data. The confusion matrix in Figure 3 shows that this method of binning increased the prediction accuracy by ~3%.



Kind of wine \ Prediction (Kind of wine)	Type1	Type2	Type3
Type1	20	0	0
Type2	5	13	6
Type3	0	1	16

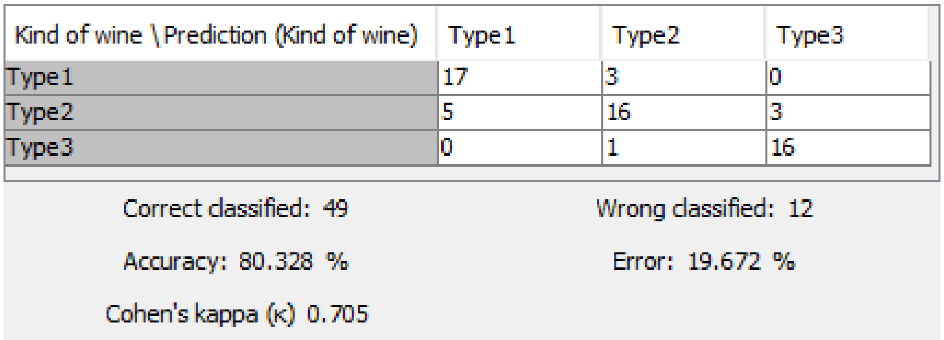
Correct classified: 49 Wrong classified: 12

Accuracy: 80.328 % Error: 19.672 %

Cohen's kappa (κ) 0.708

Figure 3: Confusion matrix of the Naive Bayes predictor using 3 manually specified bins of Flavanoids data

Figure 4 shows the confusion matrix using fixed frequency binning, which coincidentally gives the same prediction accuracy as the manually binned data, albeit with a slight difference in the misclassification between Type1 being Type2 and Type2 being Type3.



Kind of wine \ Prediction (Kind of wine)	Type1	Type2	Type3
Type1	17	3	0
Type2	5	16	3
Type3	0	1	16

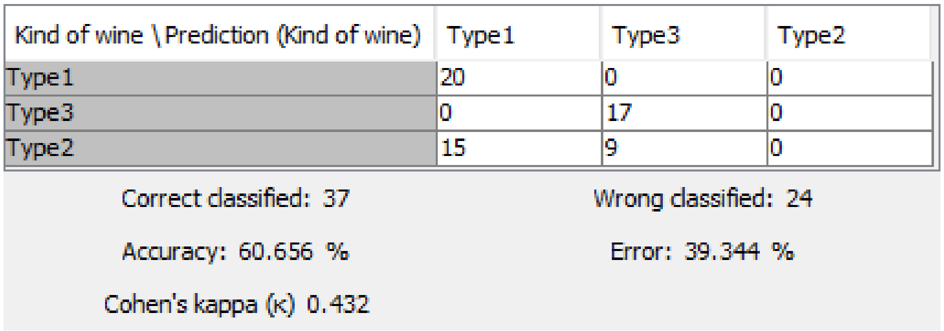
Correct classified: 49 Wrong classified: 12

Accuracy: 80.328 % Error: 19.672 %

Cohen's kappa (κ) 0.705

Figure 4: Confusion matrix of the Naive Bayes predictor using 3 fixed frequency bins of Flavanoids data

The confusion matrix displayed in Figure 5 shows that the fixed width binning method gave a prediction accuracy of ~60%, which is the worst performing binning method by a relatively large margin.



Kind of wine \ Prediction (Kind of wine)	Type1	Type3	Type2
Type1	20	0	0
Type3	0	17	0
Type2	15	9	0

Correct classified: 37

Wrong classified: 24

Accuracy: 60.656 %

Error: 39.344 %

Cohen's kappa (κ) 0.432

Figure 5: Confusion matrix of the Naive Bayes predictor using 3 fixed width bins of Flavanoids data

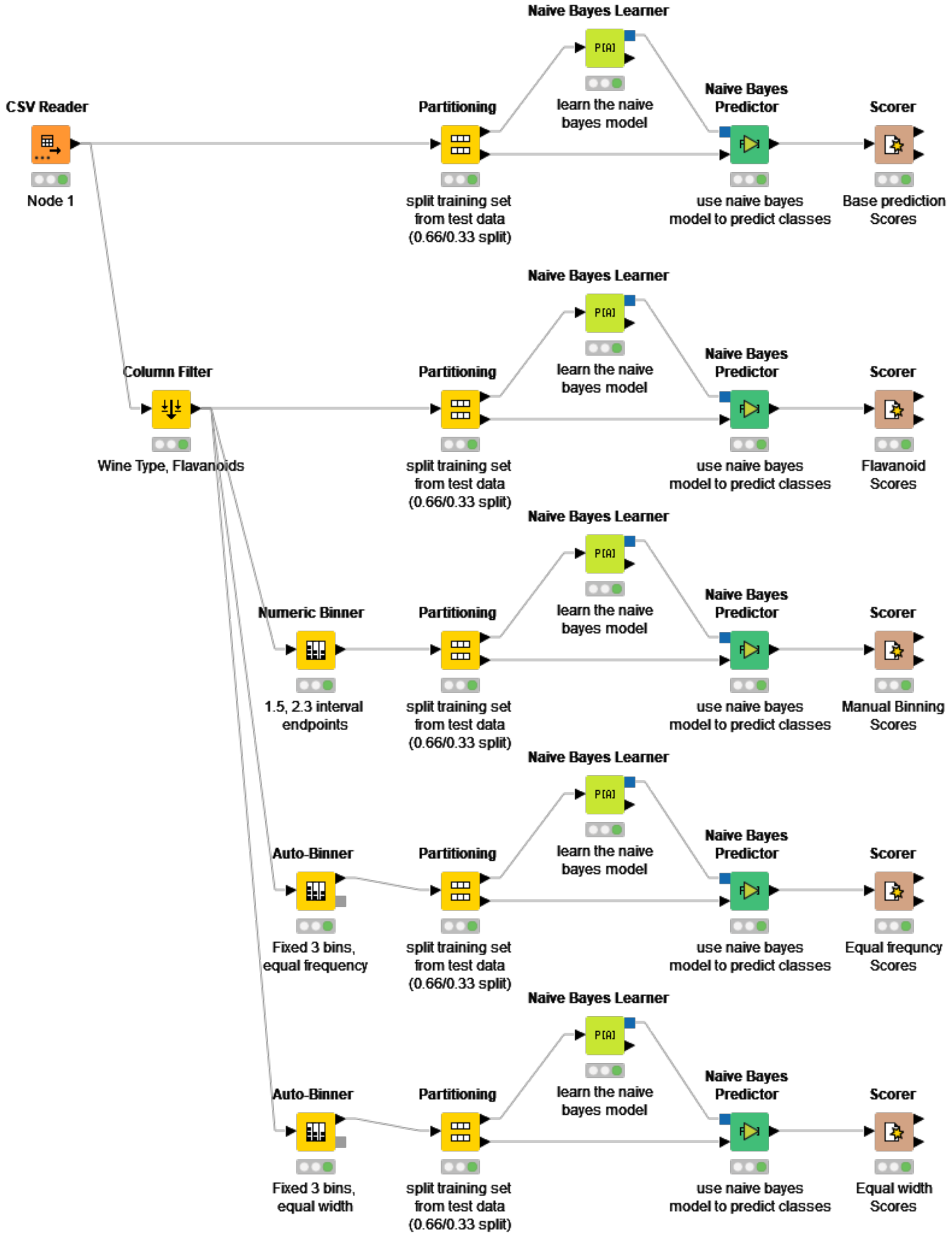


Figure 6: KNIME workflow used for analysis