1) Keep the default test parameters unchanged and set the option to 10-fold cross-validation.

a. Rank the best two methods according to the rate of correctly classified instances (% correct). Why do you think these two methods performed the best?

Without data preprocessing (e.g., discretization, standardization, normalization) the two best methods according to rate of correctly classified instances (and kappa statistic) and their scores are:

1. Naïve Bayes classification (*NaïveBayesSimple*): 173/178=97.191%
2. Nearest neighbor classifier (IBk): 169/178 = 94.9438%

I was surprised by the success of the Naïve Bayes classifier. After reading the paper we were assigned for these problems and with the knowledge of potential performance degradation with highly correlated attributes, I expected to see poorer results initially with the Naïve Bayes, finding that feature selection based on the J4.8 output led to significant improvements.

After some thought and research, I believe the success of the Naïve Bayes classifier may be a result of properties belonging to the data set and attributes working with the classifer. We know that the conditional independence assumption made by the NB classifier may degrade its performance in certain domains where feature dependencies are known, but despite this assumption and in the presence of known dependencies, the NB classifier still shows very high performance in other domains.

While correct estimation usually leads to accurate prediction, accurate prediction does not require correct estimation. It has been posited that the performance of the NB classifier is not directly correlated with the degree of feature dependencies by class, but rather how much the assumption of conditional independence results in a loss of information about the class. Further, the distribution of dependencies across each class may be important to consider because the dependencies may effectively cancel each other out.

For this data set, the output from the J4.8 tells me the attribute providing the greatest information gain for classification is Flavinoids and that Proline also provides signification information gain. The feature correlations provided tell us which variables receive more weight when predicting class. Given this information, the feature dependencies appear to facilitate the prediction capability of the NB classifier in this example.

As another probabilistic statistical learning algorithm, I would guess that these factors must also play a role in the performance level of the kNN algorithm given the redundancy from feature correlations.

b. Compare the difference between jRip and the J4.8. Are the two methods concluding the same classification rules? If not, what are the differences?

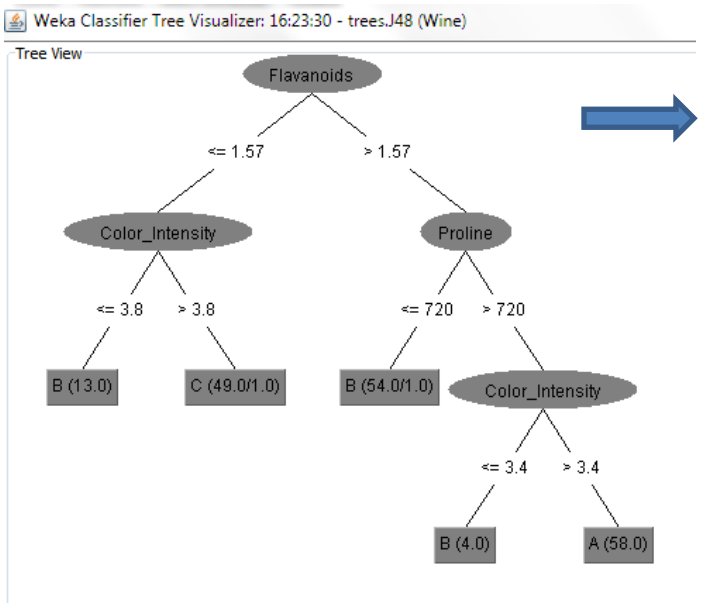
No, they are not outputting the same rules:

```
JRIP rules:
===========

(Flavanoids <= 1.39) and (Color_Intensity >= 4) => Type=C (46.0/0.0)
(OD280_OD315 <= 1.3) => Type=C (2.0/0.0)
(Proline >= 760) => Type=A (61.0/4.0)
 => Type=B (69.0/2.0)

Number of Rules : 4
```

J4.8 Tree translated to rules



(Flav <= 1.57) and (Color Int <=3.8) --> B
(Flav <=1.57) and (Color Int > 3.8) --> C

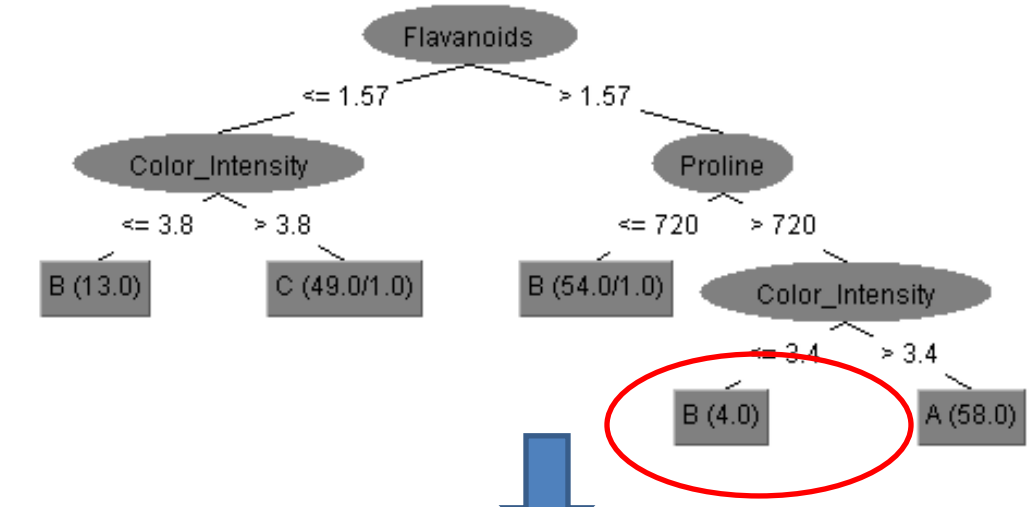
(Flav > 1.57) and (Proline <= 720) --> B
(Flav > 1.57) and (Proline > 720) and (Color Int <=3.4) -->B
(Flav >1.57) and (Proline > 720) and (Color Int > 3.4) --> A

These two classifier have similarities, but they differ in how they make decisions—jRip employs a direct method for rule extraction—without considering the entire set, the bottom-up learner beginning with the minority class label by finding a rule that explains all instances that result in that class and proceeding to the next frequent class label. The J4.8 is a top-down learner. Rules are determined indirectly; decisions are made by first considering the whole set and dividing into increasingly purer subsets based on information gain until a single class label resides among each subset of instances.

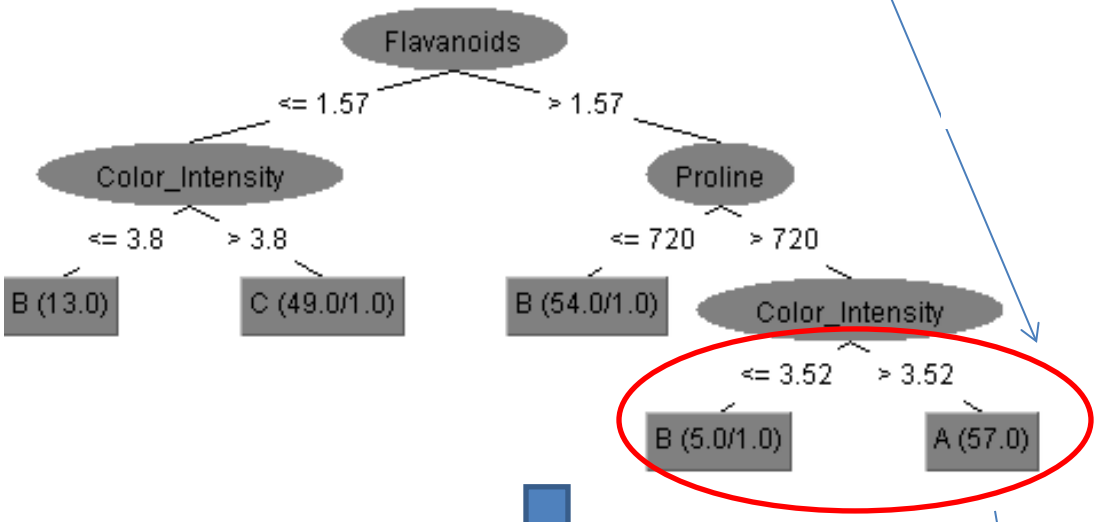
2) Now, let's examine the effect of changing the default algorithm's parameters on the analysis

*a. How does changing the "*minNumObj*" value in the* J48 *algorithm from 2 to 5 and then from 2 to 10 affect the rate of correctly classified instances and the decision tree? How does changing the "*confidenceFactor*" affect the analysis? Explain your results.*

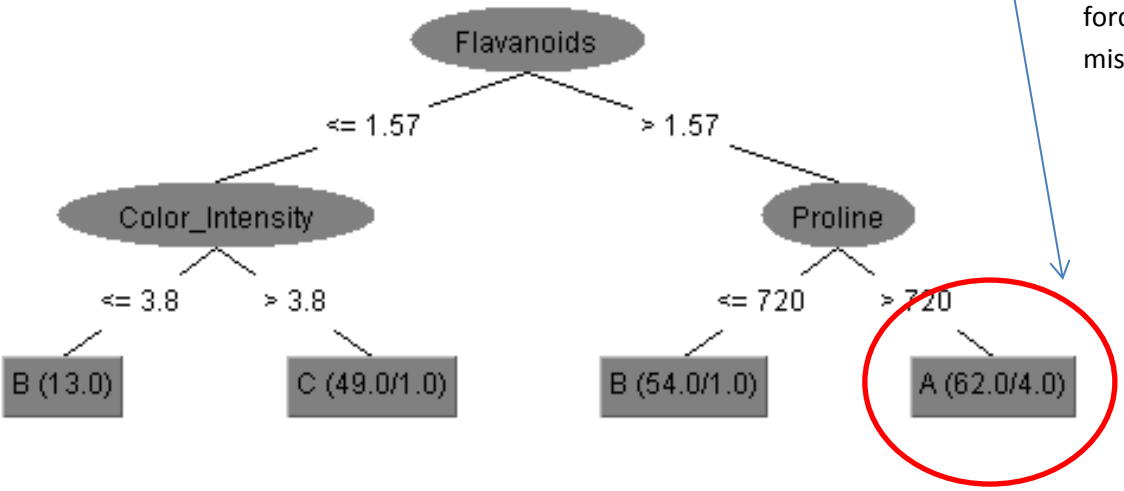
Increasing the "*minNumObj*" increases the minimum number of instances allowed per leaf and affects the tree growing behavior during online-pruning (i.e, the pruning imposed during induction). An increasing adjustment might prove beneficial with a noisy data set and helps protect against overfitting, but has the opposite effect here—increasing to 5 and then 10 results in a decrease in correctly classified instances from 93.8202% to 91.573% and 89.3258% respectively.



Increasing g the min allowed per leaf forces misclassification here



And doesn't allow for this split, forcing more misclassifications



The lower the confidence value, the more post-pruning will incur. In this case, the tree cannot be pruned any less—check the Unpruned option to True produces the same results as the original tree.

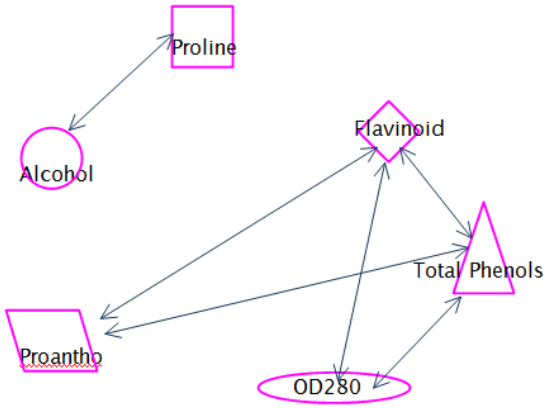
*b. How does changing the "*KNN*" Value in the* IBk *algorithm from 1 to 5 and from 1 to 15 affect the rate of correctly classified instances? Explain your results.*

Here, we see an improvement. Increasing this value changes the "k" value in the kNN—that is, it defines the number of neighboring instances to use in determining the class value. Too small a value chosen for k as well as too large a value can negatively impact the performance, due to potential for overfitting and increasing the likelihood for including "neighbors" that don't meet the rationale for the algorithm, respectively. Testing different values for optimal results is important. With the increase proposed here, we see the following improvements in correctly classified instances from 94.9438%:

5-NN: 95.5056%
10-NN: 96.0674%

c) Removing some correlated variables, do the Naive Bayes results differ? Why or why not?

I tried many combinations based on a diagram I made describing only the presence of significant correlation between attributes:



Using the rationale I describe in question 1 and knowledge of the information gain by certain attributes from the j4.8, I focused on the cluster of significant correlations. I did try removing both Alcohol and Proline separately and together, all resulted in significant reductions in classification.

Removing Flavinoids alone or in combination with anything I tried results in a decline in performance. This makes sense given its significance in the j4.8 tree in terms of information gain. I immediately though to remove Total Phenols and wanted to test removing it alone as well as in combination (but separately) with OD280_OD315 and Proanthocyanins. Despite the high level of performance we see in the presence of the correlations and its potential enhancing effect on the model in this case, the amount of correltion between this bunch in quite significant and it's possible that these benefits may still hold, while improving further by reducing unnecessary redundancy. Total Phenols appears to be the key redundancy we can remove for optimal results. Again, my explanation/rationale for choosing which to remove is described in A.

Removing only Total Phenols results in an improvement in percent correctly classified (to 97.7528%), as well as slight improvements on ROC for class A, TP rate for class B and FP rate for class C. Removing Proanthocyanine in addition to Total Phenols produced similar results, with only the ROC for A not increasing.

Refs:

Tan, P., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. Reading, MA: Addison-Wesley.
http://nlp.stanford.edu/IR-book/html/htmledition/properties-of-naive-bayes-1.html
http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf
http://www.dcc.fc.up.pt/~ines/aulas/0809/MIM/aulas/bayes08.pdf
http://wwww.samdrazin.com/classes/een548/project2report.pdf
http://www.cs.iastate.edu/~honavar/rish-bayes.pdf
http://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf
www.research.ibm.com/PM/ijcai01.ps
http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/Optimality_of_Naive_Bayes.pdf
http://masterchemoinfo.u-strasbg.fr/Documents/TutoChemo/classification.pdf
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.98.3019 (Ratanamahatana paper)
http://www.d.umn.edu/~tpederse/Pubs/pedersen-i2b2-2008.pdf