1. (1/4 point) Based on the histograms, which attribute appears to be the most useful for classifying wine, and why?

   Based on the histograms, Alcohol seems to me to be the attribute that will be most useful for classifying wine. It seems to me that it is the best histogram among them all that able the dataset into more pure labels. The splitting boundary can be drawn a little before 11.

2. (1/2 point) What is the *accuracy* - the percentage of correctly classified instances - achieved by *ZeroR* when you run it on the training set? Why is *ZeroR* a helpful baseline for interpreting the performance of other classifiers?

   The percentage of correctly classified instances achived by ZeroR on the training set is 62.381. This provides us with a benchmark to which we compare other machine learning models. One of the standard baseline performance for regression as well as classification problems is Zero Rule Algorithm. For classification, ZeroR classifies on the observation with the mode of the class values.

3. (1/2 point) Using a decision tree Weka learned over the training set, what is the most informative single feature for this task, and what is its influence on wine quality (i.e., what values of the feature are correlated with higher wine quality)? Does this match your answer from question 1?

   The most informative single feature for this task by decision tree algorithm is Alcohol. It matched with the answer from question 1.

4. (3/4 point) What is 10-fold cross-validation? What is the *main* reason for the difference between the percentage of Correctly Classified Instances when you measured accuracy on the training set itself, versus when you ran 10-fold cross-validation over the training set? Why is cross-validation important?

   Cross Validation is a technique that is used to gauge the predictive performance of a model. In order to evaluate how a model performs outside the training data, how it performs with test data. In K fold cross validation, we divide the training data in k different parts, train it on k-1 parts then test it on the kth part. This procedure is performed K times, such that each part gets to be part of the testing data.

   Results on the data:

Training Accuracy – 95.873

10 Fold Validation Accuracy – 85.9788

The reason that we get a higher accuracy for the training data is due to the fact that the model was built trying to fit the training data and now it is being tested on the same data. On the other hand, during 10 Fold Cross Validation, the model is being gauged on the validation set, which is not the same as the training set. Hence the drop in accuracy.

Cross Validation is important as it provides a more practical metric to the accuracy of the model. The validation accuracy is a more reliable metric than training accuracy to gauge a model.

5. (1/2 point) What is the "command-line" for the model you are submitting? For example, "*J48 -C 0.25 -M 2*". What is the reported accuracy for your model using 10-fold cross-validation?

Command Line for the Model Used:

RandomForest -P 100 -I 250 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Reported Accuracy: 91.164 – 10 fold cross validation

Correctly Classified Instances	1723	91.164 %
Incorrectly Classified Instances	167	8.836 %

6. (1 point) In a few sentences, describe how you chose the model you are submitting. Be sure to mention your validation strategy and whether you tried varying any of the model parameters.

I tweaked around weka and tried on multiple models, Random Forest was able to demonstrate best performance for 10-fold cross validation. I tried changing multiple model parameters, I observed an uptick in performance by increasing the number of iterations, 100 iterations led to an accuracy of 90.58, 150 iterations led to an accuracy of 91.05 and 250 iterations led to an accuracy of 91.164. Accuracy was decreasing a little after increasing the number of iterations further, hence I stopped at 250 iterations.

7. (1/2 point) Briefly explain what strategy you used to obtain the Classifiers A and B that performed well on one of the car or wine data sets, and not the other. Tell us which classifiers you chose and what their 10-fold CV accuracies were on each of the car and wine training data sets.

   After tweaking around a lot in weka tool, one of the two classifiers that suit the constraints of this question are:

   Decision Stump and Logistic Regression (10 Fold CV Accuracies)

   Decision Stump (car dataset) = 70.5%

   Logistic Regression (car dataset) = 93.5%

   Decision Stump (wine dataset) = 80.8%

   Logistic Regression (wine dataset) = 82.5%

8. (1 point) Consider the following four functions f*i* each defined over a single attribute *x*. For each data set, one of the following learners performs best in terms of leave-one-out cross validation: one-nearest-neighbor, three-nearest-neighbor, linear regression, and polynomial regression. For each function, state which of the four learning methods is best in terms of leave-one-out cross validation, and in 1-2 sentences say convincingly why this is the case. You don't have to explicitly compute the LOOCV accuracy or sum squared error in each case, although in some cases doing the computation might be a good route to a convincing explanation for why your chosen method is best.

x	f1(x)	f2(x)	f3(x)	f4(x)
1	+	3	2	+
2	+	6.5	5	+
4	-	12	17.5	+
5	-	14.5	26	-
7	+	21	49.5	+

F1 (X) - 1 Nearest Neighbor (best accuracy (80%))

F2 (X) - Linear Regression (Value of X seems to be a linear function, as x increases so does F2(x))

F3 (X) - Polynomial Regression (Non – Linear Relation between X and F3(X))

F4 (X) – 3 Nearest Neighbor (best accuracy(80%))

Regression cannot be used for classification problems, hence applied on F2 and F3.