# **Homework 5 – Handling Data with Missing Values (10 points)**

In this assignment, you will work with a data file (*cars-missing.csv*) that includes records with missing values. You will use Weka to create a J48 decision tree using the data set provided (with missing values left unaltered). You will then create another J48 decision tree after first replacing the missing values (using a filter available in Weka). Finally, you will create another J48 Decision Tree after first removing all records with missing values. You will report on the performance differences in these three circumstances.

## Directions:

* Download the *cars-missing.csv* file provided in this assignment.
* Open up *cars-missing.csv* (you need to select a loader that can read CSV files).
* Generate a J48 Decision Tree using the standard defaults.
* Copy and paste in the === Detailed Accuracy By Class === section and the === Confusion Matrix === section into your homework. Label this section: **Results with Missing Values.**
* Re-open up *cars-missing.csv*.
* Figure out how to have Weka replace the missing values in a way that does NOT take the class attribute into account. Do a little research to figure this out.
* Copy and paste in the === Detailed Accuracy By Class === section and the === Confusion Matrix === section into your homework. Label this section: **Results with Missing Values Replaced.**
* Re-open up *cars-missing.csv.*
* Figure out how to have Weka remove records (instances) with missing values. Do a little research to figure this out. HINT: The filter you need can only be applied to one attribute at a time…so look into the Multi-Filter feature. Do this step very carefully to make sure it is doing what you expect…it is easy to screw it up!
* Copy and paste in the === Detailed Accuracy By Class === section and the === Confusion Matrix === section into your homework. Label this section: **Results with Missing Values Instances Removed.**
* Indicate which scenario had the best overall precision, which had the best overall recall, and which had the best overall F-Measure. Include your answers in your submission.

## **BONUS (4 Points):**

* Convert the *cars-missing.csv* CSV file into an ARFF file (easy to do in Weka!)
* Create a Java program that will do the following:
  + Read in the ARFF file
  + Generate the same output as done for the **Results with Missing Values Replaced** section. Be sure to label this section.
  + Generate the same output as done for the **Results with Missing Values Instances Removed** section. Be sure to label this section.
* Turn in the Java program and include a snapshot of the output.

If you don’t do the bonus, just turn in a Word file or TEXT file called *homework5.doc* or *homework5.txt*. If you DO the homework, zip up all of your files into one file called *homework5.zip*.

**Warning**: This is an individual homework and should be done by yourself. Observe the university integrity policies as mentioned in the course syllabus. Late submissions turned in prior to class will receive a 20% penalty. Submissions turned in after class begins will not be accepted.