**Part 1: Visually explore the data (6 pts) For each dataset, describe.**

*1. The types of each attribute*

***Dataset file***

**a.arff**

**Graphical user interface, application

Description automatically generated**

TYPE **NUMERIC** FOR FILE **a.arff**

**b.arff**

A screenshot of a computer

Description automatically generated with medium confidence

TYPE **NUMERIC** FOR FILE **B.arff**

**c.arff**

**Graphical user interface, application

Description automatically generated**

TYPE **NOMINAL** FOR FILE **c.arff**

*2. The distribution of the class labels*

**Distribution of the class labels for** **a.arff**

*Graphical user interface

Description automatically generated*

**Distribution of the class labels for** **b.arff**

**Graphical user interface

Description automatically generated**

**Distribution of the class labels for** **c.arff**

**Graphical user interface, application

Description automatically generated**

*3. Any noticeable qualitative characteristics*

***Trends are statistical patterns that appear over time or across different values of a variable. Finding trends may aid in understanding how data changes and can be used to make predictions or forecast future behaviour.***

***Outliers are data values that deviate dramatically from the remainder of the data set. These numbers can have a considerable influence on measures of central tendency and dispersion, and they should be checked to see if they are real or if they indicate data mistakes.***

**Part 2: Algorithm Evaluation (8 pts each)**

● Use 10-fold cross validation to test/compare Decisions ump and J48 on dataset c. Here are the steps to do this: ○ Select the explore button.

■ Click on the open file button and select the “c.arff” file.

■ With the pre-process tab open, make sure the "Class (Nom)" field is selected on the right side of the screen (it should be by default). On the left, select different attributes (click the names, not the check boxes) and observe the distributions of data over the two classes (1 and 2). Specifically note the distribution of the "class" attribute. You can also click on "visualize all" to look at all distributions at once.

**A picture containing chart

Description automatically generated**

Now click on the classify tab to switch to it

■ Click on the choose button, and then select: Trees > DecisionStump

■ Make sure “(Nom) Class” is selected as the attribute to predict (it should be by default).

■ Under test options, make sure Cross-validation is selected and Folds is set to 10 (default).

■ Click start, and review the output on the right.

Graphical user interface, text, application

Description automatically generated

■ Repeat these steps for the classifier

s ● Trees > J48 ● Trees > J48 (unpruned) (to set the option for unpruned trees, click on J48, then click on the text bar to the right with various flags/settings. This should bring up a window with many different options, one of which is a dropdown box labeled “unpruned” with options for true and false. Set this to true to use unpruned decision trees)

Graphical user interface, text, application, email

Description automatically generated

■ In a table, list the classification accuracy (correctly classified instance percentage) and the RMSE for each classifier. (One row for DecisionStump, two rows for J48(pruned/unpruned.)

Graphical user interface, text, application, email

Description automatically generated

■ For DecisionStump, briefly explain the technique and list the attribute that was used to make the decision. Compare the results of J48(pruned/unpruned) and explain why pruned has better performance. Hint: compare the depth of the created decision trees in both cases, and consider how depth is related to complexity and generalization.

Graphical user interface, text, application, email

Description automatically generated ● Run J48(pruned), Naive Bayes, k-NN (IBk) with k=1 and k=21 (the “k” parameter is set the same way you set the “unpruned” parameter), on data sets a.arff and b.arff leaving the other parameters with their default values ○ For each classifier, use (weighted average) F-measure to compare its performance obtained on data set a to its performance obtained on data set b. ○ For data set a, compare the performance of the 4 classifiers using F- measure. ○ Explanation why you think you got the results you did.

Graphical user interface, text, application

Description automatically generated

● Run Naive Bayes, IBk (k=1 and k=10) on data set c using default parameters ○ Compare the performance of the 3 classifiers using (weighted average) F-measure. Comment on the effect of the parameter “k” ○ Explanation why you think you got the results you did. ● Hint: if weka has trouble computing the weighted average F-measure automatically, you might have to compute this by hand from the confusion matrix.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated