WQD7006
Machine Learning for Data Science

Test – 15%
I Hour

Please save your submission file using your name.

No screen shots allowed, unless stated otherwise.

Answers must be precise according to the questions.

Part 1: Use the vote dataset (part of Weka dataset)

1. In ONE sentence, describe what the dataset is about? (0.5 mark)

2. Briefly describe the dataset using descriptive statistics. Focus on the MAIN points only. (1 mark)

3. Run Naïve Bayes using all the attributes with a 10-fold cross validation.

   3.1 Report the classifier's performance for democrat and republican. (2 marks)

   3.2 Report the overall classifier's performance for 3.1 above. (1 mark)

   3.3 Draw the confusion matrix for 3.1. Explain (1 mark)

   3.4 Run Naïve Bayes again, however, use 80 - 20% split. Compare the performance with (3.1) in terms of accuracy. What can you conclude? (2 marks)

   3.5 Differentiate 10-fold cross validation and a split-data. Use diagram to aid your explanation (2 marks)

4. Use the same dataset, 10-fold cross validation. Run Logistic Regression.

   4.1 Report the overall classifier's performance. (1 mark)
   4.2 How do you think it performed compared to Naïve Bayes? (1 mark)

5. Use the same dataset, 10-fold cross validation. Run J48
   5.1 What is the size of the tree generated? (0.5 mark)

   5.2 Draw the decision tree. You may use the screen shot for this question (1 mark)

   5.3 Report J48's classification performance. (1 mark)

6. Compare Naïve Bayes, Logistic Regression and J48, and report if they are significantly different in classifcation. Use the default setting. Report the necessary results. (1 mark)

THE END