

National Institute of Technology Calicut
Department of Computer Science and Engineering

Winter 2019-2020

CS4038D Data Mining – Assignment 1

Total marks: 20 Due Date: 28/03/2020 8PM

Instructions:

1: Original work by an individual is expected to be submitted.

2: Any form of plagiarism if found may lead to zero marks.

3: You must create a single PDF file containing answers to the following questions.

4: Your entire source codes in Python or any other tools along with the PDF file must be submitted as a single compressed file in the CSE Eduserver.

1: Download any classification dataset from the website:

<https://www.kaggle.com/tags/classification>

2: Answer to the following questions based on your dataset.

1. How many attributes are there in the dataset? How many are Nominal, Ordinal and Numeric?
2. How many records in your dataset?
3. Are there any missing values? If yes, provide details like how many such records with missing values, which attribute has more number of missing values?
4. Did you apply any data cleaning process in the dataset to improve its quality? Justify your answer.
5. Give your problem statement (describe what is your aim in implementing a data mining model for this dataset).

3: Divide your dataset into train and test set using the 80:20 method. Apply Decision tree and Naïve Bayesian classification algorithms on the dataset.

1. Report the classification Accuracy, Error rate, Sensitivity, and Specificity. Display the confusion matrix.
2. Is there class imbalance problem with your dataset? Justify your answer.
3. Which classification algorithm in question 3 performs better with your dataset?

4: For decision tree algorithm, try with three different attribute selection methods and report which performs better in your problem.

5: Plot the final tree obtained, with highest accuracy based on question 4, for your problem.