Lab 4 – SECTION A , BATCH 1 Date:29 th Nov 2021

Exer 1: Use the Cereals.csv data set and WEKA to answer the following questions:

Perform the following preprocessing steps

1. Replace missing value with mode or means.
2. Convert vitamins, shelf to nominal attribute.
3. Design a Linear regression model to predict the rating of a cereal based on all nutrients. Tabulate the accuracy of the model using
    a. 80 % , 20 % split
    b. Using cross validation.
4. Use select attributes tab and CfsSubsetEval to identify the top 5 related nutrients.
5. Design a Linear regression model to predict the rating of a cereal based on top 5 related nutrients. Tabulate the accuracy of the model using
    a. 80 % , 20 % split
    b. Using cross validation.
6. Discretise rating into 5 bins using the filters.
7. Design a Logistic regression model to predict the rating of a cereal based on top 5 related nutrients. Tabulate the accuracy of the model using
    a. 80 % , 20 % split
    b. Using cross validation.

Exer 2: Using the SUPERMARKET.ARFF data set and WEKA to answer the following questions:

1.Split the dataset into 2 datasets , 1 containing items and the other containing departments.

2.For the item data set, find the 5 most frequent itemsets ranked as per support.

3. Which are the top 5 selling items in the dataset ?

4. For the top selling items, find association rules with the item on the RHS of the rule. Tabulate the support, confidence and lift of the rule.

5. Find top 5 association rules for the department store. Tabulate the support, confidence and lift of the rule.