

### **ARTIFICIAL INTELLIGENCE ASSIGNMENT 03**

This assignment is not code-oriented. It is designed to give you an exposure of a new tool where models are present. You just need to incorporate them into your dataset in order to produce desired results. The tool to be used is Rapidminer. You will submit question#.properties and question# .rmp file

#### **NAÏVE BAYES CLASSIFIER**

QUESTION 1) Naïve Bayes classifier is a famous approach for supervised learning. It mainly classifies a test data provided with the fact that training data is used to train up the model. Within the folder, winequality-red dataset is to be used for this question. There exist 13 features and 1 label named as pH.

- a) pH is the class label which needs to be predicted. As the testing data is not separately provided thereby, you will have to split this dataset for training and testing respectively. Use the ratio of 70:30 for training and testing respectively.
- b) Train Naïve Bayes model using 70% of the dataset and then classify the rest 30% of the data.
- c) Measure performance parameters i.e. accuracy, precision and recall to show how much accurate the model has been for the dataset.
- d) There are two different models within rapidminer for naïve bayes; use both of them separately and check the accuracy difference.

Question 2) Use the same naïve bayes model on a new dataset. Parkinson training and testing data now shall be used for this question. Within this question we no more need to split data. First training data shall be used to train the model and then testing data shall be used to find out its labels.

- a) Parkinson training data has 22 features and 1 label column. The last column represents the class label. Currently it has the values of 0 or 1. However naïve bayes within rapidminer will throw an error if we used this training data set because it does not accept numerical class label. Thereby load the data and then create one more attribute named as classes which basically converts this numerical class label values into nominal ones. After creating this attribute as label, you now may delete the last column of the data because now you have simply replaced that column with another column.
- b) Train naïve bayes model using the training dataset.
- c) Repeat (part a) on Parkinson testing dataset now.
- d) Apply it on model and check performance of the model as it is the most important step. The performance measure should show accuracy, precision and recall.

- e) There are two different models within rapidminer for naïve bayes; use both of them separately and check the accuracy difference.

### **DECISION TREES CLASSIFIER**

Question 3) Apply decision tree model on question 2 and report the accuracy difference.

Question 4) Use adult.csv dataset now to use ID3 algorithm of decision trees. Income class attribute needs to be used as label. As training and testing data is not separate thereby split the data into 60:40 ratios for training and testing respectively. Apply the model and then check its performance i.e. accuracy, precision and recall.

### **K- MEANS CLUSTERING**

Question 5) K-means clustering has not been taught in the class thereby this question involves your study about clustering. You are not required to study the entire working on how clustering algorithm works. You just need to understand what k-means clustering does to any particular dataset and how its validation is done.

- a) Use the winequality-red dataset. As we are provided with label of pH thereby we have ground truth. Make pH as label and remove attribute citric from the features.
- b) Apply k-means clustering and then compare the results of clusters obtained with ground truth.
- c) Check performance of the model as it is the most important step. The performance measure should show accuracy, precision and recall.
- d) Use alternate performance measure of silhouette coefficient. (Study first that what this measure does and why it is used).