AI ASSIGNMENT 03
DUE DATE: SUNDAY, 10TH MAY 2020, 11:50 P.M.

Question 1) Classify the data set using Gaussian Naïve Bayes Classifier.

Dataset description:

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voices.

The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

The task involves the implementation of the classifier to an interesting problem of speech classification for Parkinson's disease.

I. You are required to implement each and every step of Gaussian naïve bayes classifier from scratch in order to classify this dataset. Data has been divided into 80/20 ratio of training and test data set.
II. Evaluate the performance of the classifier in the form of accuracy. This means that the ending part within your code should be to match your result of testing data with the ground truth of testing data in order to show up that how good your classifier is.

NO LANGUAGE RESTRICTION

The following questions are 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 – download it and read its tutorials as it's a simple drag and drop tool which enables you to fit different models on your dataset**.

You will submit question#.properties and question# .rmp file for this question.

NAÏVE BAYES CLASSIFIER

Question 2) 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, wine 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. Take snapshots of these results and add it into a word document.

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

Question 3) 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. Take snapshots of these results and add it into a word document.
e) There are two different models within rapidminer for naïve bayes; use both of them separately and check the accuracy difference.

K- MEANS CLUSTERING

Question 4) K-means clustering is an unsupervised learning approach. 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 wine 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. Take snapshots of these results and add it into a word document.