PGCP DSML 08

Module: Machine Learning

Assignment-2 (real dataset)

---

1) Read the breast cancer dataset from sklearn as per the following details:
   (from sklearn.datasets import load_breast_cancer)
   i) Check the data by converting the built-in data set to a Pandas dataframe.
   ii) Split the data into training and test sets (a ratio of 75:25)
   iii) Build a logistic regression model on the above data
   iv) Plot the ROC and calculate AUC for the training data and find the best threshold.
   v) Based on this threshold, classify your test data.
   vi) Also check the balancing of the dataset in terms of benign and malignant classes. Make use of oversampling technique to balance the dataset and then check the classification accuracy on the oversampled dataset.

2) Extract the stock market data from Yahoo finance for a set of any 15 companies of four different sectors during 01.04.2017 to 31.03.2020. Make use of the following attributes on the data:
   - Highest price of the day
   - Minimum price of the day
   - Opening price of the day
   - Closing Price of the day

   Take the average of these four attributes as an average movement of the day. Now apply k-Means clustering to cluster the chosen 15 companies as per their respective sectors. Write a brief summary (50-70 words) of your clustering results.

3) Explore the application of k-means clustering in color compression of a color image. To explore it, read an image having multiple colors where a large number of colors will be unused, and many of the pixels in the image will have similar or even identical colors. Cluster these several colors into a 16 different clusters and show the compressed image.