DS514- Data Mining for Business Applications
Assignment 1

Date: 1st Mar ‘24 **Total Marks: 100**

Deadline: 11th Mar ‘24, 11:55 PM

Instructions:

1. The assignment should be done individually.
2. You should submit all your work in **one** Jupyter Notebook. The notebook should follow naming convention:
   <entry_no.>_DS514_Lab1.ipynb
   If you are using additional dataset for analysis, you should include them in a .ZIP file. It should follow the same naming convention:
   <entry_no.>_DS514_Lab1.zip
3. This is a high level description of the problems. In addition to given tasks, you are free to use your own creative analysis and draw insights from the data. Kindly make sure to support your work with relevant facts.
4. You can find the dataset link in the description of the problem.

Q.1. Use the taxi trip dataset to find clusters of locations that might belong to a particular cluster having similar kinds of trip features. Use DBSCAN algo for this.

- Show the locations belonging to each cluster over a map.
- Perform the analysis of trips belonging to the same clusters and point out the distinguishing features and insights.
- Also, divide a day in 1-hour slot and show a time of day based analysis of each cluster

Note: To map the ZoneId(Pickup and Dropoff) to their spatial coordinates, use the Zone Lookup table from the link below. This will be used to create visualization of clusters on the map.

Dataset link:
https://drive.google.com/file/d/1L97I_30HY-Q4BKujiPWPODjkvPo2VnIm/view?usp=drive_link

Zone Lookup table:
https://drive.google.com/file/d/1tIVIoCyGKRaVsxN_KOS3T5O-SIZPrvqp/view?usp=sharing

Refer to the document for working with PARQUET format.
You can find more information about the dataset at:
https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

[40 Marks]

Q.2. Use the city-wise vehicle registration dataset of million+ population cities of India.

- Perform K-means and Agglomerative clustering this dataset. Discuss the insights and distinguishing features of the cities belonging to the same cluster.
- If you need to design a marketing campaign which cities you will target for what kind of vehicles.
- Use data.gov.in to check the other kinds of datasets available that can help you with your analysis. Substantiate your results with proper facts and figures.

Dataset link:
https://drive.google.com/file/d/1FXOk4TAeKepHDXivUx95KDjbXQ0wx5v7/view?usp=drive_link

[40 Marks]

Q.3. Use Spotify dataset for Association rule mining.

- Identify user playlist patterns based on features like Artist name, Playlist name etc.
- Find association rules based on frequently played artists/playlists. Determine which of them classify as strong.

Dataset link:
https://drive.google.com/file/d/1Lj3xGBwmzqbUhLVdfcqHftI3sPk4_NI-/view?usp=sharing
[20 Marks]