**Assignment 1**

**Data Analytics using Clustering**

*Fundamental of Big Data Spring 2023*

**Due Date:9th March 2023**

In this assignment, you have to pre-process and cluster the dataset provided to you **using Python. You have to submit your python code along with a word document explaining and analyzing your results and findings.**

1. Pre-Process the dataset using different techniques like normalization, discretization, and correlation. Explain which pre-processing steps are performed on the data and **why?**
2. Select a subset of attributes for Clustering*.* You can select multiple subsets as well
3. Perform **K-MEANS** clustering on the given dataset.
   1. Use the Elbow method to find K and run K-means with **different K**
   2. For each value of K, run K-means multiple times. Report your findings (error in each Clustering, the time required)
   3. Report the K that gives the best result for each subset of attributes.
   4. Indicate the number of iterations to convergence for different runs.
   5. Examine the **quality of clusters** and also of **clusterings**. Report the errors: within-cluster sum of squared error, between-cluster sum of the square error, and silhouette coefficient for each run of K-mean.
4. Also, Cluster the dataset using **Hierarchical Clustering (single link, complete link, average link) and DBSCAN.** 
   1. Run for different values of the number of clusters. Include the dendrogram(for hierarchical) and time taken by each Clustering in your report.
   2. Examine the quality of clusterings. Report the errors.
5. Draw different plots to visualize the clustering results (include plots in your report).
6. ***Compare the clustering results of the K-means, Hierarchical, and DBSCAN in terms of time and quality of Clustering.***