**Machine Learning I : Capstone Project**

1) A proposal for the project (one page long) (DUE 4/14/2024)

**About Dataset**

This dataset is designed for learning customer segmentation concepts, such as market basket analysis. It includes basic customer data such as Customer ID, age, gender, annual income, and spending score, which is assigned based on customer behavior and purchasing data. The goal is to help a supermarket mall owner understand their customers better, identify target customers who are likely to converge, and provide insights to the marketing team for strategic planning.

**Description of the problem.**

The primary challenge faced by the supermarket mall is the lack of a targeted marketing strategy that addresses the diverse needs and preferences of its customer base. Without a clear understanding of the different customer segments, the mall risks deploying generic marketing campaigns that fail to engage potential customers effectively.

Kaggle Link: https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmentation-data/data

**Description of the dataset (dimensions, names of variables with their description)**

* Customer ID: A unique identifier for each customer.
* Age: The age of the customer.
* Gender: The gender of the customer (Male/Female).
* Annual Income (k$): The customer's annual income in thousands of dollars.
* Spending Score (1-100): A score assigned to the customer based on their behavior and purchasing data.

**Description of the techniques that you will use**..

I will apply EDA, Data Preprocessing, K-Means Clustering, Hierarchical Clustering, Principal Analysis Component Analysis (PCA), silhouette score.

**Comments and/ or concerns?**

As with most data, it would have been better if there were more features, but because it’s simple, it’s better for me to try algorithms with my understanding. Data might be flawed considering the usability score of 9.41.

This project includes the following deliverables:

1) A proposal for the project (one page long) **(DUE 4/14/2024)**

* Description of the problem.
* Description of the dataset (dimensions, names of variables with their description)
* Description of the techniques that you will use..
* Comments and/ or concerns?

2) A presentation **(DUE 05/02/2024 & 05/09/2024)**

* Description of the data and the question/s that you are interested in answering
* Review of some of the approaches that you tried or thought about trying.
* Summary of the final approach you used and why you chose that approach.
* Summary of the results.
* Conclusions.

In preparing the presentation you should be aiming it at a smart audience. Hence, you should not just say “I did Supervised Learning” but also explain the basic idea of how it works, why it might be better than other etc. Among other things, points will be allocated for clear articulations of the question of interest, the approach you used to solve it, the reason you chose that approach, and the conclusions you were able to draw.

3) A report **(DUE 05/11/2024)**

* The report will contain a summary of the material covered in the presentation (maximum 3 pages).
* The first page should be an executive summary.
* The report must also include the slides from the presentation and a technical appendix, which should include your code from Jupyter