Machine Learning

CAI26332 Project:

1. Introduction

This project is designed to give students the opportunity to apply their machine learning skills to a real-world problem. Students will learn how to collect, prepare, analyze, and present data. They will also learn how to work as team and how to deliver their results and findings to audience.

2. Your project will show that you can.

- Collect dataset from different source.
- Preprocessing data for machine learning models.
- Apply machine learning models on your dataset.

3. Project Requirements:

1) Groups Formation

This project should be completed by group of 3 students or less.

2) Data Collection

1. Each group should choose a data source and collect a dataset of at least 400 data samples. These can be structured or unstructured data sources such as Google map Reviews, YouTube comments, Hotel /restaurant reviews, Airbnb etc. If you have to use web scrappers /crawlers be sure to adhere to ethic of web scrapping.

2. Avoid use Kaggle dataset or any other prepared dataset.

4. Data Analysis.

Students should use analytical methods and algorithms to analyze the data and extract insights.

- - ***Exploratory data analysis(EDA)*** : perform descriptive statistics, graphical representations to gather insights about sample distribution, trend and

patterns on your collected data from which Algorithms can leverage on during modeling.

- - ***Modeling***: Depending on the nature of collected data and the targeted problem apply at least three supervised ML or Unsupervised or approaches learned in class to analyze your data. Here are possible choices:
    i) *Supervised learning algorithms***:** linear regression, logistic regression, K- nearest neighbor (KNN), decision trees, Random Forest, Extreme Gradient boosting (xgboosts) etc.

    ii) *Unsupervised learning algorithms*: Principal component analysis (PCA), Clustering algorithms, Isolation Forest etc.

5. You will submit.

- Presentation including all the project details, explains the process of each step in collecting and preprocessing the dataset. [50%].

- Your final files of preprocessing (.ipynb) and dataset (.CSV). [20%].

- Students should create a 10-20mins PowerPoint presentation to deliver their findings and insights from the analysis. Group members will meet, discuss and review the PowerPoint presentations before presenting it to seminar audience [30%](each student will be evaluating individually based on this answers to some questions from students and instructor).