# Introduction to Data Science

# Semester Project

**(Due Date: 06 June 2024)**

**Objective:**

The goal of this assignment is to demonstrate proficiency in conducting Exploratory data analysis (EDA) using Python, specifically Pandas and NumPy, and to create insightful visualizations and finally apply Machine Learning Algorithms. Students are expected to work with real-world datasets from Kaggle to derivemeaningful insights and communicate them effectively through visualizations.

**Requirements:**

A. **Dataset Selection:**

**To select a dataset for your project, please follow these instructions:**

1. Visit the Kaggle website: https://www.kaggle.com/
2. Browse through the available datasets and choose one that meets the following criteria:
   - Contains at least 10 columns.
   - Has 5000 or more records.
   - Is relevant to your interests or field of study.
3. Once you have selected a dataset, send the following information to the class representative (CR):
   - Dataset name
   - Link to the dataset on Kaggle
4. The class representative will acknowledge your dataset selection and ensure that it has not been chosen by any other student.
5. Dataset selection is on a first-come, first-served basis, so make sure to select your dataset promptly.

B. **Data Cleaning and Preprocessing:**

- Handle missing data using appropriate imputation strategies.
- Detect and address outliers or anomaliesin the dataset.
- Perform any necessary feature engineering or transformation.

C. **Exploratory Data Analysis (EDA):**

- Utilize Pandas andNumPy for in-depth data analysis.
- Conductstatistical analysis, including measures of central tendency, dispersion, and correlation. • Perform advanced grouping andaggregation, exploring relationships between multiple variables. • Extract meaningful insightsfrom the data through descriptive statistics.

D. **Data Visualization:**

- Create a variety of visualizations using Matplotlib and Seaborn.
- Include histograms, box plots, scatter plots, and correlationmatrices.
- Implement interactive visualizationsif applicable, using toolslike Plotly. (Optional)
- Clearly annotate and label visualizationsfor effective communication.

E. **Hypothesis Testing:**

- Formulate at least two hypotheses related to the dataset.
- Use statistical tests(t-test, chi-square, etc.) to validate or refute the hypotheses.

F. **Advanced Analysis:**

- Apply at-least two machine learning techniques relevant to the dataset.
- Validatemodel performance and interpret results.

G. **Documentation:**

- Provide a detailed Jupyter Notebook documenting the entire analysis process.

• Include explanationsfor each step, rationale behind decisions, and interpretations of results. • Use Markdown cells for clear and concise explanations.