COM2502 Introduction to Data Science Course Project Assignment

1. Project Overview

The course project is designed to apply the fundamental concepts of data science to a real-world dataset. Students will go through the complete data science pipeline, including data collection, cleaning, exploratory analysis, visualization, and modeling (if applicable). The project should focus on deriving insights from data and presenting findings effectively.

2. Project Objectives

By completing this project, students will:

- Gain hands-on experience working with data.
- Develop skills in data preprocessing and visualization.
- Apply machine learning techniques for analysis.
- Communicate insights clearly through a structured report.
- Work with tools such as Jupyter Notebook, Google Colab, and use libraries such as Pandas, and Matplotlib.

3. Project Requirements

A. Dataset Selection

- Students must use a **publicly available dataset** or collect their own data.
- The dataset should contain at least **1,000 rows** and multiple features (columns).
- Sources for datasets: Kaggle, Hugging Face, UCI Machine Learning Repository, Google Dataset Search, etc.

B. Data Preprocessing & Analysis

- **Data Cleaning:** Handle missing values, duplicates, and outliers.
- **Exploratory Data Analysis (EDA):** Generate summary statistics and visualizations.
- **Feature Engineering:** Create meaningful features (if applicable).
- **Data Visualization:** Use appropriate charts/graphs to present insights.

C. Classification or Regression Problem

- Apply **at least three machine learning models for a classification or regression problem**.
- Justify the choice of the method and evaluate its performance.
- Use metrics such as accuracy, precision, recall, F1-Score, RMSE, etc. (if applicable).

D. Tools & Technologies

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn (if applicable) etc.
- **Notebooks:** Use Jupyter Notebook or Google Colab, or work on your local computer with an IDE.

4. Project Deliverables

A. Project Proposal (in PDF format) – [10% of the grade]

1. **Title Page** – Project title, group number, student name(s), student id(s), course, date.
2. **Introduction** – Problem statement, objectives, and dataset description.
3. **Methodology** – The models will be used.
4. **References** - Cite dataset sources and any external materials used.

Note that I will be giving feedback to each group after project proposal submission.

B. Project Report (in PDF format) – [50% of the grade]

The report should be well-structured and include:

1. **Title Page** – Project title, group number, student name(s), student id(s), course, date.
2. **Introduction** – Problem statement, objectives, and dataset description.
3. **Methodology** – Data preprocessing steps, analysis techniques, and models used.
4. **Results & Findings** – Visualizations, insights, comparison of the models and key observations.
5. **Conclusion** – Summary, limitations, and future improvements.
6. **References** – Cite dataset sources and any external materials used.

C. Project Source Code Submission (ZIP file) – [20% of the grade]

- The source code should be well-structured and commented for readability.
- Create a ZIP file containing all your project source code files. (The source code files can be Python files (.py) or Notebooks (.ipynb).)

D. Presentation (5-10 minutes) – [20% of the grade]

- Present findings, insights, and key takeaways.
- Note that the presentation cannot exceed 10 minutes.
- Demonstrate your project using a single sample to perform classification or regression.
- Use slides (PowerPoint, Google Slides) to support the explanation.
- Be prepared to answer questions from the instructor and peers.

5. Submission Guidelines

- **Project Proposal Submission:** Submit your project proposal in PDF format until **03/04/2025, 11:59 PM**. Name your project proposal as `GroupNumber_ProjectProposal.pdf.`
- **Project Deadline:** Submit project report and source code ZIP file until **31/05/2025 11:59 PM**.
- **File Naming Convention and Submission:** Create a ZIP file containing all your project source code files. Name the ZIP file as `GroupNumber_Project.zip`. Upload it to e-campus. Also name the project report as `GroupNumber_ProjectReport.pdf.`
- You are going to upload ZIP file and report seperately in E-campus.

- Please do not forget to include your dataset link in your project report. If there is no link exists, include the dataset in the ZIP file.
- **Late Submission Policy:** Late submissions will incur a penalty of **10% per day**.

6. Special Considerations

- Project Group Selection will be opened to choose your groups.
- The recommended group size is **2**. **If you can’t find any group mate, you can work alone.** Note that even if you choose to work alone, you must still select a group in the Project Group Selection.
- Ensure ethical data handling and avoid bias in analysis.
- Bonus points (5 points) **for using extra machine learning algorithms**, **interactive dashboards, web app integration, or innovative approaches**.