

Introduction to Data Science Project Instructions

You are required to complete the following project as part of the **Introduction to Data Science** course. Please read the instructions carefully:

Objective

1. Select a **unique dataset** (no two students can work on the same dataset).
 2. Perform **Exploratory Data Analysis (EDA)** on the dataset with **at least 10 to 15 different analyses**.
 3. Preprocess the dataset for further analysis.
 4. Apply any **machine learning model** of your choice to the preprocessed dataset.
 5. Use **Streamlit** to create an interactive web application to present your project.
-

Step-by-Step Guide

1. Dataset Selection

- Choose a dataset that is interesting and has enough features and records to perform meaningful analysis.
- Add your dataset name and link in the provided excel sheet. To avoid duplication, you can verify your dataset from a sheet. ***First-come, first-served!***

2. Exploratory Data Analysis (EDA)

Perform the following (and other relevant) analyses:

- Summary statistics (mean, median, mode, etc.)
- Visualizations (e.g., histograms, scatter plots, box plots)
- Correlation analysis
- Missing value analysis
- Outlier detection
- Feature distribution analysis
- Data types and unique value counts
- Trend analysis (if applicable)
- Grouped aggregations (e.g., by category or time)
- Insights from relationships between features (pairwise analysis)
- Use any additional analyses relevant to your dataset.

3. Data Preprocessing

Prepare the dataset for machine learning:

- Handle missing values (imputation or removal).
- Encode categorical variables.
- Scale or normalize numerical features.
- Split the dataset into training and testing sets.
- Perform any other necessary transformations.

4. Machine Learning Model

- Choose any supervised or unsupervised machine learning model based on your dataset (e.g., regression, classification).
- Train the model and evaluate its performance using appropriate metrics.

5. Streamlit Application

- Build an interactive application using **Streamlit** to showcase your project.
- Include the following sections:
 - **Introduction:** Briefly describe your dataset and project objectives.
 - **EDA Section:** Display visualizations, tables, and key insights from your analysis.
 - **Model Section:** Show your model, predictions, and evaluation results.
 - **Conclusion:** Highlight the main takeaways from your project.
- Ensure the application is user-friendly and visually appealing.

Deadline for choosing Dataset: 21 Dec, 2024

PS: This is an individual project.

Note: If you need any help at any stage, feel free to ask questions.

Good luck!