



Final Project

Project Overview:

In this project, teams of **3-5** members will collaborate to explore, analyze, and model a dataset using Linear Regression to estimate a value.

Tasks:

1. Dataset:

Find a dataset containing more than **10,000** records.

2. Exploratory Data Analysis (EDA):

Conduct a thorough EDA to uncover patterns, anomalies, trends, and relationships within the data. Visualizations should be used to help understand the distribution of data and the relationships between features.

3. Data Cleaning:

This should cover issues like missing values, outliers, and inaccurate data entries.

4. Model Development:

Build a Linear Regression model. The model should be robust, and its parameters should be fine-tuned to get optimal performance. Evaluate the model using appropriate metrics.

5. Regression Diagnostics:

Perform Regression Diagnostics to validate the model assumptions.

Important Guidelines:

- Make sure to split the data into training and testing datasets **BEFORE** anything else to avoid data leakage and ensure model generalization.
- Avoid dropping records unless it's extremely necessary and this should be well documented and justified.
- You're required to provide the complete model pipeline, from data preprocessing to final evaluation.

Bonus:

Deploy the trained model using a framework such as Flask, FastAPI, or Streamlit.

Create a simple web-based user interface (UI) that allows users to:

- Upload or input data
- Receive model predictions

Discussion:

ALL team members should be present in the discussion which will be held on **May 22nd**.

You should have a notebook prepared that includes:

- i. An overview of the dataset, explaining the types and nature of features.
- ii. Insights and visualizations from the EDA.
- iii. Dimensionality reduction technique(s) used.
- iv. Modeling pipeline.
- v. Hyperparameter tuning and model evaluation.