Final Project

Topics

- English Premier League (EPL) Match Data 2000-2025
  - https://www.kaggle.com/datasets/marcohuiii/english-premier-league-epl-match-data-2000-2025
- Credit Scoring in Banks
  - https://www.kaggle.com/competitions/GiveMeSomeCredit/overview
- Length-of-Stay Prediction in Hospitals
  - https://www.kaggle.com/datasets/nehaprabhavalkar/av-healthcare-analytics-ii
- Insurance Premium Analysis
  - https://www.kaggle.com/datasets/schran/insurance-premium-prediction

Final Project

Requirements

- Deliverable: Formatted notebook (.ipynb) with all performed steps, visualizations, and explanations. (Upload to Moodle before the deadline).
- The presentation:
  - The presentation should be done with slides/keynotes, **NOT** the submitted notebook.
  - Time: 3-5 minutes per member.

Final Project

Requirements – Presentation (20%)

- The presentation should have at least the following main parts:
  - Short introduction about the dataset.
  - One to two problem statements / research questions to be answered using the dataset.
  - Cleaning and preparing the data.
  - Generating summary statistics.
  - Exploring and visualizing the data.
  - Testing assumptions and providing potential answers to the problem statement(s).
  - Optional: Deploying suitable machine learning algorithm(s) according to your objective.
- Note: All steps above should be done with the problem statements / research questions in mind.

Final Project

Requirements – Notebook (20%)

- The submitted notebook should include:
  - **Readability.**
    - Clear and concise description of the **task** and **research question**.
    - Modular code design using Python **functions** (e.g., def data_imputation()).
    - Include **docstring** and **instructions** for all major code components.
  - **Completeness**.
    - Fully functional code that covers **all steps** show in your presentation.
    - **Visualization** of key results (e.g., plots of model performance, error trends, feature importance).
  - **Reproducibility**
    - Ensure that the notebook can be run from start to finish without errors on a new environment.
      - Use **fixed random seeds**.
      - Avoid **hardcoded paths** and environment-specific configurations.