

650-9 - Introduction to Machine Learning

Machine Learning Mini-Project

In this project, you will apply the concepts learned in the course to a real-world dataset. The goal is to explore, preprocess, and model the data, ultimately presenting your findings and insights in a structured report.

Project Description

You will analyze a real-world dataset of your choice (e.g., from [Kaggle](#)) to address a specific business or analytical question. The dataset should be suitable for either regression or classification tasks. Your objective is to:

- Preprocess the data for analysis.
- Explore and visualize relationships within the dataset.
- Train and evaluate machine learning models to answer the question.
- Compare results from different modeling approaches and interpret their implications.

Project Requirements

1. Dataset Selection

- Choose a dataset that is rich enough to allow meaningful analysis and modeling.
- Provide a brief description of the dataset and the question it addresses.
- If you are searching for a dataset on Kaggle, you can follow these steps:
 - Explore Kaggle's "[Competitions](#)" or "[Datasets](#)" sections and use keywords related to your topic of interest.
 - Filter datasets by popularity or recency to find well-documented and widely used datasets.
 - Ensure the dataset you select has sufficient features and observations for your analysis.

2. Data Exploration and Preprocessing

- Perform exploratory data analysis to understand the structure, relationships, and distributions of your data.
- Handle missing values, outliers, or other anomalies in the dataset.
- Transform or scale features as necessary.

3. Modeling

- Train multiple machine learning models to address the chosen question. Use either regression or classification approaches as applicable.
- Compare different algorithms and assess their performance using appropriate metrics.
- Optionally include techniques like cross-validation or hyperparameter tuning.

4. Insights and Recommendations

- Interpret the results of your models in the context of the question.
- Compare the performance of the different methods and explain the trade-offs involved.
- Provide actionable insights or recommendations based on your findings.

5. Reporting

- Prepare a concise, structured report summarizing your analysis, methodology, results, and conclusions.
- Include visualizations to support your findings.
- The report should not exceed 5 pages (excluding visualizations and appendices).

Evaluation Criteria

- **Problem Understanding and Dataset Selection (20%):** Relevance and quality of the dataset and clarity of the question.
- **Exploratory Data Analysis and Preprocessing (20%):** Thoroughness and correctness of data preparation steps.
- **Modeling and Evaluation (30%):** Quality of models, evaluation, and comparison of results.
- **Insights and Reporting (20%):** Depth and clarity of insights, along with report quality.
- **Innovation (10%):** Creative use of methods or novel approaches in the analysis.

Submission Instructions

- Submit your code (Jupyter Notebook or Python script) and the final report as a PDF.
- Ensure all code is well-documented and reproducible.
- Provide a link to the dataset or include it in your submission if permissible.