Machine Learning Final Project

Instructor: Dr. Nayel Bettache

Fall 2024

Project Overview

The final project for this course will be a take-home data analysis assignment. Students are required to select a dataset from Kaggle that they enjoy, and the dataset should be suitable for performing regression and/or classification. Specific questions for analysis will be provided to guide the project. Avoid considering datasets that are already studied on publicly available Github repositories.

The project requires students to work in groups of 3 to 5 members. Students must finalize and have their groups approved by the instructor by November 21st. Any student who has not joined a group by this deadline will be assigned to a group by the instructor. The final report documenting the results of your analysis must be submitted by December 23rd.

Project Requirements

- **Dataset:** The project requires you to choose a dataset from Kaggle. The dataset should be of sufficient size and complexity to perform both regression and/or classification tasks. You are encouraged to choose a dataset that aligns with your personal interests. See this project as an opportunity to build a portfolio you will be able to present during interviews.
- **Group Formation:** Students will work in groups of 3 to 5. Groups must be finalized and approved by the instructor no later than November 21st. Any student who has not joined a group by this date will be assigned to a group.
- **Final Report:** The report should be no longer than 8 pages (excluding the code). It must be submitted as a PDF by December 23rd. Late submissions will incur a 20% penalty if submitted within 24 hours after the deadline; reports submitted later than 24 hours will not be accepted. The report should clearly present your findings, explain the methods used, and provide accurate interpretations of the results.
- **Analysis Scripts:** All data analysis must be conducted using either R or Python, and the scripts used in your analysis should be submitted alongside the final report. The code will not count towards the 8-page limit but must be fully reproducible.

Grading Criteria

Your project will be graded based on the following criteria:

- **Application of Methods:** Effective and correct application of the methods discussed throughout the course (e.g., data preprocessing, feature selection, model building and evaluation).
- **Clarity and Organization:** The report should be well-organized and easy to follow. The analysis should be clearly explained, and the results should be presented logically.
- **Interpretation of Results:** You should provide meaningful interpretations of your findings, explaining the implications of your analysis.
- **Reproducibility:** Your analysis should be fully reproducible using the provided scripts. This means that the dataset, code, and analysis steps must be clearly presented and accessible to others.
- **Originality:** The creativity and originality of your approach to solving the problem will be assessed. Innovative techniques, thoughtful experimentation, and new insights into the dataset will enhance your project score.

Project Scope

This project offers an opportunity for students to apply the knowledge and skills developed throughout the semester in a practical and meaningful way. It allows you to explore real-world data and demonstrate your ability to perform complex data analysis tasks, from data preprocessing to modeling and evaluation.