Course Project

This is a group project. Each group will have 3 people. This project is marked out of 60 points.

Due Date: February 6 Friday, 2026, 11:59PM

Project Overview

In this project, you will explore the use of supervised machine learning methods for a classification task. You should choose three out of the five methods (i.e., Logistic Regression, K-Nearest Neighbors, Linear Discriminant Analysis, Decision Trees, and Random Forests) to predict the survival of passengers aboard the Titanic using the provided augmented dataset. The objective is to build models that are both efficient (using a minimal number of features) and highly performant.

The survival of passengers is the target variable, with Survived indicating whether a passenger survived (1) or not (0). The dataset includes passenger demographics, family relationships, ticket and cabin details, as well as some additional attributes. The dataset is provided alongside this document for use in the project.

To accomplish the project, you should perform the following procedures:

- Data preprocessing: Use appropriate techniques to preprocess data (e.g., normalization, standardization).
- Data splitting and resampling: Split the dataset into training and test sets (25% test) and use appropriate resampling techniques (e.g., k-fold cross-validation) to ensure robust model evaluation.
- Model building and training: Build three proper models with the three supervised machine learning methods chosen and train them on the dataset. Consider feature importance to reduce the number of features while maintaining performance.
- Hyperparameter tuning: Tune the respective hyperparameters of the chosen models to improve their performance.
- Result evaluation and visualization: Evaluate model results with specific metrics (loss and accuracy if applicable) and visualize them in an intuitive way (e.g., loss/accuracy vs. epoch graphs if applicable).
- Analysis and discussion: Analyze and discuss the performance of the chosen models and the strategies you used for performance improvement, as well as the difficulties you encountered and how you solved them.

Notes:

- For full instructions and detailed dataset description, please refer to the accompanying README file.
- The program should be written in Python, and it is allowed to use machine learning libraries such as Scikit-Learn.

Deliverables

The deliverables are (1) a project report, (2) an .ipynb file, and (3) a project presentation.

The project report is submitted as a single PDF file (12-point font and double-spaced). It should include the following sections:

- Abstract (150-250 words): A concise and factual summary of the project conducted.
- Introduction (1 page): A brief overview of background and context to convey the importance of supervised machine learning methods in classification tasks.
- Methodology (2-3 pages): A detailed description of how the project was conducted.
- Experiment (2-4 pages): A detailed description of experimental design, results, analysis, and discussion.
- Conclusion (0.5 page): A general summary of the main ideas and insights in the project conducted.
- References ($\leq$ 1 page): A list of the sources of the cited information.
- Contributions of group members (if applicable).

The .ipynb file contains the scripts (including the results after running) used for result replication. It should be accessible, understandable, and fully reproducible.

The project presentation highlights and summarizes the project conducted using slides. It should be planned for 8 to 10 minutes.

Marking Rubric

In this project, you will be evaluated based on the following criteria, with weightings and details:

Criteria	Weightings	Details
The validity of the methodology and experiment	50%	(1) Justification for the choice of data preprocessing techniques, data resampling approaches, model building and training strategies, and hyperparameter tuning strategies (2) Thorough and thoughtful analysis and discussion of how your choices affect the performance of the models and how you solved the difficulties encountered
The performance of the models	30%	(1) How well the models perform in the classification task according to the metrics used (2) Why were the specific features selected for training the model, and how do they contribute to the performance of the models
The clarity and conciseness of the project	20%	(1) Conveying information with clear and unambiguous expression (2) Avoiding unnecessary details and lengthy explanations

report and presentation		

To earn bonus marks, you are encouraged to apply state-of-the-art deep learning methods that were proposed after 2022 (e.g., Transformers). The bonus method should be an add-on to the original methods mentioned in the project description.

Expected Outcome

At the end of this project, you are expected to have gained theoretical knowledge and hands-on experience in building and training supervised machine learning models for classification tasks. You will develop a solid understanding of how key hyperparameters influence model performance and how to tune them effectively. Additionally, you will appreciate that not all features contribute equally to model predictions, and you will have experience selecting the most relevant features to create efficient, high-performing models. Good luck with your project!