Course Project

This is a group project. Each group will have 2-3 people. This project is marked out of 60 points. If you plan to do it individually, please let the instructor know.

Due Date: February 9 Friday, 2024, 11:59PM

Project Overview

In this project, we will explore the use of supervised machine learning methods for a classification task. You should choose two out of the five methods (i.e., Logistic Regression, K-Nearest Neighbors, Linear Discriminant Analysis, Decision Trees, and Random Forests) to classify the quality of milk based on its seven features (i.e., pH, Temperature, Taste, Odor, Fat, Turbidity, and Colour).

The quality of milk is classified into three grades: low (bad), medium (moderate), and high (good). pH, Temperature, and Colour are quantitative features, and they are given actual values. Taste, Odor, Fat, and Turbidity are qualitative features, and they are assigned 1 (satisfactory) or 0 (unsatisfactory). The dataset and more information can be found at https://www.kaggle.com/datasets/cpluzshrijayan/milkquality/data.

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

- Data preprocessing: Use appropriate techniques to preprocess data (e.g., normalization, standardization).
- Data resampling: Use appropriate approaches to resample data (e.g., leave-one-out cross-validation, k-fold cross-validation).
- Model building and training: Build two proper models with the two supervised machine learning methods chosen and train them on the dataset.
- Hyperparameter tuning: Tune the respective hyperparameters of the two 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 (loss/accuracy vs. epoch graphs if applicable).
- Analysis and discussion: Analyze and discuss the performance of the two models and the strategies you used for performance improvement, as well as the difficulties you encountered and how you solved them.

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) a .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 (100-150 words): A concise and factual summary of the project conducted.
- Introduction (0.5 page): A brief overview of background and context to convey the importance of supervised machine learning methods in classification tasks.
- Methodology (1-2 pages): A detailed description of how the project was conducted.
- Experiment (1-2 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 (≤ 1 page): A list of the source 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 replicable.

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
The clarity and conciseness of the project report and presentation	20%	(1) Conveying information with clear and unambiguous expression (2) Avoiding unnecessary details and lengthy explanations

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 the project, you expect to have gained theoretical knowledge and hands-on experience in building and training supervised machine learning methods for classification tasks, and have had a good understanding of the key hyperparameters that affect the performance of the models and how to tune them. Good luck with your project!