2104618 Project Assignment for Second Semester of Academic Year of 2567

Problem Statement

The problem concerns water safety. You are asked to develop classification models to predict whether the water is safe based on the provided features.

The dataset contains twenty features, which represent the composition of water, and a label indicating its safety status as follows:

Features (Attributes)

- aluminium
- ammonia
- arsenic
- barium
- cadmium
- chloramine
- chromium
- copper
- fluoride
- bacteria
- viruses
- lead
- nitrates
- nitrites
- mercury
- perchlorate
- radium
- selenium
- silver
- uranium

Further apply correlation analysis and cut the weak correlation

Predict variable (desired target)

- is_safe – 1 if the water is safe, 0 if the water is not safe

Instruction

- ❖ The dataset contains 7,996 examples and is randomly divided into three parts: training data (4,796 samples – 60%), cross-validation data (1,600 samples – 20%), and test data (1,600 samples – 20%).
- ❖ The name of each column can be found in the provided MS Excel file.
- ❖ Develop forecasting models to predict whether the water is safe using the training data. Then, compare the models using the cross-validation data to select the best one. Finally, determine the generalization error of the selected model using the test data.

- You may use any supervised learning algorithm, such as logistic regression (with or without regularization), neural network, support vector machines, random forests, decision trees, gradient boosting, XGBoost, LSTMs, etc.
- You must develop **at least three algorithms** for comparison. Within each algorithm, you may create multiple models by varying the hyperparameters or structures.
- The more algorithms you use, the higher your score will be.
- You do not have to use all features; you may select a subset of the provided features if preferred.
- You should explore the features during the preprocessing step. If there are any missing values, you may impute them (e.g., replacing them with the mean, median, mode, previous value, or next value).
- You may use polynomial features and add interaction terms if necessary.
- You may use any algorithm to train the models, such as the scipy.optimize module or gradient descent.
- You may use any Python library.
- Feel free to set hyperparameter values yourself, such as the regularization parameter, learning rate, threshold, number of hidden layers, and number of neurons per hidden layer. Please clearly state the hyperparameters and their values for each model.
- You may use GridSearchCV, RandomSearchCV, or BayesSearchCV to find the optimal combination of hyperparameters.
- If you use neural network, you may choose any activation function (e.g., ReLU, ELU, PReLU, Sigmoid, tanh, etc.).
- Performing machine learning diagnostics, such as checking bias/variance, plotting learning curves, or debugging, is recommended.
- You may evaluate your models using any criterion, such as classification accuracy, classification error, F1 score, precision, or recall. Please justify your choice of evaluation metric.
- Prepare a report and submit it on the presentation days:
  - Thursday, May 1, 2025 (for Section 1)
  - Sunday, May 11, 2025 (for Section 9)

  In addition, send the Jupyter Notebook files for all models to the instructor's email: nantachai.k@chula.ac.th.

- The report should include the following sections:
  1. Table of Contents
  2. Table of Tables
  3. Table of Figures
  4. Problem Statement

5. Literature Review
6. Methodology
7. Results and Discussion
8. References

- Do not forget to cite references in the Literature Review section.
- Prepare a presentation. Each presentation will last 15 minutes, followed by a 5-minute Q&A session.
- **Honor Code: You agree to complete this project independently without external assistance.**