IDATG2208 - Mandatory Assignment - 2

Deadline for submission - 10 Oct, 2025

Instructions

Download the **Breast Cancer Wisconsin dataset** to explore Decision Trees and Support Vector Machines (SVMs) for this assignment. The dataset is available directly in `scikit-learn` (`https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html`) and also on Kaggle (`https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data`). All the exercises should be solved by splitting the dataset into **60% training, 20% validation, and 20% test**. Use the training set for **5-fold cross-validation**, the validation set for **model selection and hyperparameter tuning**, and reserve the test set strictly for **final evaluation**. For each experiment, report the **mean** and **standard deviation** of accuracy (and, where specified, other metrics such as precision, recall, or AUC) among the folds. Your submission should consist of a concise report that includes code, results in the form of tables or figures, and answers to the questions provided below.
The answers should be submitted through a pdf document on Blackboard. The answers should be supplemented with the figures and code snippets. Alternatively, solutions can be provided on Github (link should be provide on Blackboard).
Each question is mandatory and needs to be answered. A minimum of 70 points is needed for the solution to qualify as valid submission for this mandatory assignment.

Exercise-1: Data Preparation [10 points]

Q1.1 Load the dataset, inspect feature names and target distribution. Comment on dataset imbalance. [3 pts]

Q1.2 Analyze all features with and without standardization (i.e., zero mean and unit variance). Plot the feature analysis with and without standardization and decide which version is more suitable. [3 pts]

Q1.3 Comment on importance of three way split with respect to hyperparameter search and robustness of any learned model. [4 pts]

Exercise-2: Decision Trees [30 points]

Q2.1 Train a Decision Tree classifier using default parameters. Evaluate it on validation sets from original splits (report accuracy mean and std). [5 pts]

Q2.2 From the trained model, comment on feature importance values and identify the top 3 features from your model. [5 pts]

Q2.3 Vary the `max_depth` parameter (e.g., depth 2-10). Use validation accuracy (mean ± std from cross-validation on the training set) to choose the best depth. Provide performance for each chosen depth (at-least 5 to be reported) and discuss the aspects of overfitting vs. underfitting. [10 pts]

Q2.4 Repeat previous exercise with different `min_samples_leaf` values. Which setting generalizes best according to the validation set? [10 pts]

Exercise-3: Support Vector Machines (SVM) [30 points]

Q3.1 Train a linear SVM (`kernel="linear"`) and evaluate on validation sets (use the original 5 splits that was created in previous exercise). Report accuracy mean, std and plot the ROC for each split. [5 pts]

Q3.2 Train an Radial Basis Function (RBF) kernel SVM. Compare its performance to the linear kernel using validation accuracy. Plot ROC for these models. [5 pts]

Q3.3 Experiment with different values of C (regularization strength). Use the validation set to select the best C. Report results as a plot of accuracy vs. C. [10 pts]

Q3.4 Experiment with different γ values for the RBF kernel. Discuss the effect on bias-variance trade-off for all experimented values. Select the best γ using the validation set and report the performance on validation set. [10 pts]

Exercise-4: Model Comparison [30 points]

Q4.1 Compare Decision Tree and SVM results from the training set (cross-validation mean ± std) and validation set. Plot the performance comparison plots (e.g., scatter plots, ROC curves). Which model generalizes better? [5 pts]

Q4.2 Discuss the trade-off between usability and accuracy for this dataset. Which model would you recommend for a medical decision-support system, and why? (Hint - Make use of F_β score analysis) [10 pts]

Q4.3 Use `GridSearchCV` with the training set to tune hyperparameters for both Decision Trees and SVMs. Confirm your final choice with the validation set. Summarize the best settings and provide relevant performance plots. [5 pts]

Q4.4 Compare the final test set accuracy of the best Decision Tree and best SVM. Which model performs better in practice? [5 pts]

Q4.5 The Breast Cancer Wisconsin dataset has 30 continuous features, many of which are correlated and not linearly separable. Explain why a linear SVM might fail to capture complex patterns in this dataset. How does using an RBF kernel help in this case? Discuss your answer in terms of the dataset's feature space and the geometry of the decision boundary. [5 pts]