Course Project Instructions

1 Overview of the Project

For each dataset, students must conduct experiments in two phases:

- **Phase 1 (Baseline):** Train classifiers using the original feature set (no feature selection or dimensionality reduction).
- **Phase 2 (After FS/DR):** Apply the assigned feature selection or dimensionality reduction method, then train the same classifiers using the reduced feature set.

For each phase, evaluate:

- 5 classifiers
- 10-fold cross-validation
- 2 evaluation metrics: Accuracy and F1-score (macro-F1 for multi-class)

Total evaluations per dataset:

$$10 \text{ folds} \times 5 \text{ classifiers}$$

2 Excel Sheets Structure

Two Excel sheets must be completed:

- Sheet 1: **Before FS/DR (Baseline)**
- Sheet 2: **After FS/DR**

Both sheets must be filled independently.

Do not change or remove any row/column of the excel sheet.

Columns correspond to classifiers and rows correspond to cross-validation folds and metrics.

3 Cross-Validation Procedure

For each dataset and classifier:

1. Use 10-fold cross-validation.
2. Split the original training dataset into training and validation folds.
3. Perform hyperparameter tuning using only the training dataset; the test set must not be used.
4. Train the model on the training fold.
5. Evaluate on the validation fold.
6. Record results.

Important: Data leakage is strictly prohibited.

The following operations must be performed inside each fold only:

- feature scaling
- feature selection or dimensionality reduction
- hyperparameter tuning

Feature selection must **not** be applied to the entire dataset before cross-validation.

4 Recording Results in Excel

For each fold, record:

- Accuracy
- F1-score
- Parameters
- Number of selected features (for FS/DR)
- The selected features name (for FS only)
- FS/DR parameters (for FS-DR)

Rules:

- Enter one value per cell.
- Use decimal format (e.g., 0.8732).
- Do not use percentages.
- Do not change the template excel file (no addition, no removal), otherwise penalties may apply.

5 Summary Statistics (Mandatory)

At the bottom of each classifier column compute:

- Mean
- Standard deviation

Report results as:

$$\text{mean} \pm \text{std}$$

Example:

$$0.912 \pm 0.021$$

6 Hyperparameter Tuning

Hyperparameters must be tuned using cross-validation (nested cross-validation). The training data must be split into training and validation sets, with 10% used for validation.

- Grid search in the inner loop
- Training data only
- Cross-validation inside the training folds

Model performance must be estimated using an outer cross-validation loop. The test set must be used only for the final evaluation.

7 Recommended Hyperparameters to Tune

- **Support Vector Machine:** C, kernel, gamma
- **k-Nearest Neighbors:** number of neighbors (k), distance metric
- **Decision Tree:** max depth, min samples split
- **Random Forest:** number of estimators, max depth
- **Multilayer Perceptron:** hidden layer sizes, learning rate, regularization (α)

8 Reporting Hyperparameters in the Report

Students must report the best hyperparameters for each classifier and each phase.

Example Table

Classifier	Phase	Best Hyperparameters
SVM	Before	$C = 10$, kernel=rbf, gamma=0.01
SVM	After	$C = 5$, kernel=rbf, gamma=0.001
kNN	Before	$k = 7$
kNN	After	$k = 5$

Table 1: Best hyperparameters selected via cross-validation

9 Reporting Final Results

Students must report final results using mean $\pm$ standard deviation.

Example Format

Dataset	Phase	SVM	kNN	DT	RF
Data 1	Before	0.89 ± 0.02	...	...	...
Data 1	After	0.93 ± 0.01	...	...	...

Table 2: Classification performance comparison

10 Discussion Requirements

Students must discuss:

- best performing classifier
- performance differences between phases
- impact of feature selection/dimensionality reduction
- feature reduction amount
- trends and observations
- statistical improvements if applicable

11 Common Mistakes to Avoid

- performing feature selection before cross-validation
- reporting only averages without standard deviation
- using test data during tuning
- mixing folds across models
- manual entry errors

12 Submission Checklist

Students must submit:

1. completed Excel sheets (Before and After)
2. written report (Word or LaTeX) including:
   - 2.1 hyperparameter tables
   - 2.2 mean $\pm$ std results
   - 2.3 figures and analysis
3. source code