Assignment 1 (Individual Assignment: Supervised Learning – Coding) - Due

Nov 2, 2025 11:59 PM

Fall 2025 Machine Learning (DAMO-640-10)

Assignment 1 (Individual Assignment: Supervised Learning – Coding)

Implement a full ML pipeline in Python using Modules 1–4 on the Haberman's Survival dataset (binary classification).

Dataset URL:

https://archive.ics.uci.edu/dataset/43/haberman+s+survival

Column Names:

1. age (in years)
2. operation_year (year of operation, 1958–1969)
3. axillary_nodes (# of positive lymph nodes detected)
4. survival_status (1 = survived ≥ 5 years, 2 = died within 5 years)

Your **Jupyter notebook** must perform each of the following steps:

1. Data Loading & EDA (Module 1)

- Load the CSV (no header) and assign the four column names.
- Display first five rows, summary statistics, and the class distribution (survival_status).

- Check for missing or invalid values.

2. Preprocessing (Modules 1 & 3)

- Convert survival_status to 0/1 (0 = no, 1 = yes).
- Split into 75 % train / 25 % test (random_state=42).
- Standardize the three feature columns (age, operation_year, axillary_nodes) using StandardScaler.

3. Dimensionality Reduction (Module 4)

- Apply PCA on the training features.
- Plot cumulative explained variance vs. number of components.
- Choose the smallest number of components that retain ≥ 90 % of variance and transform both train and test sets.

4. Supervised Learning (Module 2)

- Train Logistic Regression and Decision Tree classifiers on the PCA-transformed training data.
- For each model, compare two hyperparameter settings:
    - Logistic Regression: C = [0.1, 1.0]
    - Decision Tree: max_depth = [3, None]
- Use 5-fold cross-validation on the training set to compute mean accuracy for each setting.

5. Model Evaluation & Optimization (Module 3)

- On the test set (PCA-transformed), for each model's best hyperparameter setting, compute:
    - Accuracy, precision, recall, F_1-score
    - ROC curve and AUC
- Plot both models' ROC curves on the same axes and discuss which is preferable and why.

6. Brief Report

In a Markdown cell or separate PDF (≤ 2 pages), summarize:

- Key EDA findings (class balance, feature distributions)
- PCA results (chosen components and variance retained)
- Cross-validation accuracies for each hyperparameter setting
- Test-set metrics and ROC comparison
- Final recommendation

Submission

The following needs to be submitted:

- Notebook (.ipynb), reproducible top to bottom.
- Report (embedded Markdown or separate PDF, ≤ 2 pages).

Submit both (or a single .zip) via the LMS by the deadline.

□

Assignment 1 (Individual Assignment: Supervised Learning – Coding)

□ Due November 2 at 11:59 PM

□ Starts Oct 19, 2025 12:01 AMEnds Nov 2, 2025 11:59 PM