

**Assignment 1 (Individual Assignment: Supervised Learning – Coding)**

Prumucena, Fabio (NF1002000)

Fall 2025 Machine Learning (DAMO-640-10)

University of Niagara Falls

Professor:  [Ali El-Sharif](https://www.linkedin.com/in/elsharifali/overlay/about-this-profile/?lipi=urn%3Ali%3Apage%3Ad_flagship3_profile_view_base%3BKqVftQQ%2BSRW0fojkglZceA%3D%3D)

**CONTENT**

[1. Introduction 3](#_Toc212473655)

[2. Key EDA Findings 3](#_Toc212473656)

[3. Data Processing 3](#_Toc212473657)

[4. PCA – Principal Component Analysis (Dimensionality Reduction) 4](#_Toc212473658)

[5. Supervised Learning and Hyperparameter Tuning 4](#_Toc212473659)

[6. Test Evaluation and ROC Analysis 4](#_Toc212473660)

[7. Recommendation 4](#_Toc212473661)

[8. Reproducibility Notes 5](#_Toc212473662)

# Introduction

This study uses the *Haberman’s Survival Dataset* (UCI) containing post-surgery survival outcomes of breast-cancer patients.  
The objective is to build a supervised learning pipeline—aligned with Modules 1 to 4 of the course—to predict five-year survival as a binary classification.  
The dataset comprises four variables: **age**, **operation year** (1958–1969), **axillary nodes** (number of positive lymph nodes), and **survival status** (1 = survived ≥ 5 years, 2 = died < 5 years).

# Key EDA Findings

The exploratory analysis showed class imbalance in the target variable. *Survival status* was recoded to 0/1 (1 → 1, 2 → 0) for consistent metric evaluation.  
Feature distributions revealed that **age** clustered around middle adulthood, **operation year** remained within 1958–1969, and **axillary nodes** were right-skewed with a few high values.  
No missing data were found, and checks confirmed that all operation years and lymph-node counts were within valid ranges.

# Data Processing

The preprocessing stage ensured clean and comparable inputs.The target variable was encoded as {1: 1, 2: 0}, and the data were split into **75 % training** and **25 % testing** sets using random\_state = 42 with stratified sampling to maintain class balance. Numerical features—**age**, **operation year**, and **axillary nodes**—were standardized with **StandardScaler**, fitted on the training data and applied to the test set.

# PCA – Principal Component Analysis (Dimensionality Reduction)

Dimensionality reduction was performed using **PCA** on the standardized training data.  
The **cumulative explained variance (CEV)** curve was plotted to determine component retention.  
The smallest number of components preserving at least **90 %** of total variance was identified with np.searchsorted (0.90) and applied to both training and test sets, yielding a compact yet informative feature representation.

# Supervised Learning and Hyperparameter Tuning

Two classifiers—**Logistic Regression (LR)** and **Decision Tree (DT)**—were trained on the PCA-transformed training data and, in accordance with the assignment guidelines, hyperparameters were tuned as follows: *LR:* C ∈ {0.1, 1.0}; *DT:* max\_depth ∈ {3, None}.  
A **five-fold cross-validation** using accuracy as the scoring metric provided mean and standard deviation values to support model comparison and selection.

# Test Evaluation and ROC Analysis

Both models were retrained with their optimal hyperparameters and tested on the PCA-transformed data. Performance metrics—**accuracy**, **precision**, **recall**, and **F₁-score**—were computed, and **ROC curves** were plotted on the same axes. **Logistic Regression** achieved a higher AUC and more balanced metrics, indicating stronger generalization and stability through regularization.

# Recommendation

**Logistic Regression** is recommended as the final model due to its superior AUC, interpretability, and robustness against overfitting—qualities especially valuable in small medical datasets.

# Reproducibility Notes

The notebook executes fully and reproducibly, using fixed random\_state = 42 and explicit hyperparameter grids. All rubric requirements—EDA, scaling, PCA, CV tables, test metrics, and ROC analysis—are implemented in the [appendix](https://github.com/prumucena1979/personalMLTerm4) for transparent verification.