Individual Assignment: Supervised Learning – Coding

Student Name: Fabio dos Santos Prumucena (NF1002000) **Course:** Advanced Data Analytics (DAMO-640-10) **Institution:** University of Niagara Falls Canada **Instructor:** Ali El-Sharif **Date:** October 28, 2025

1. Introduction

This report details the implementation of a supervised learning pipeline on the Haberman’ s Survival Dataset, which contains post-surgery survival outcomes for breast cancer patients. The objective is to build and evaluate classification models to predict five-year survival as a binary outcome. The methodology adheres strictly to the assignment requirements, covering data preparation, dimensionality reduction, hyperparameter tuning, and comparative model evaluation.

2. Key Findings and Data Preparation

Exploratory Data Analysis (EDA)

The initial EDA confirmed data integrity, with no missing values detected across the three features: patient age, year of operation, and the number of positive axillary nodes. The analysis revealed a significant **class imbalance** in the target variable, where 73.5% of patients survived ≥ 5 years (Class 1) and 26.5% died within 5 years (Class 0). This imbalance was managed by using stratified sampling during the data split.

Preprocessing and Dimensionality Reduction (PCA)

The data was split into a 75% training set and a 25% test set using stratified sampling and a fixed random state (42) for reproducibility. All features were standardized using

`StandardScaler`.

Principal Component Analysis (PCA) was applied to the standardized training data to reduce the feature space. The analysis showed that **two principal components (PCs)** were sufficient to retain 90.2% of the total variance in the dataset. Consequently, the 3-dimensional feature space was successfully reduced to 2 dimensions for all subsequent model training and evaluation.

3. Supervised Learning and Model Evaluation

Two classification models, Logistic Regression (LR) and Decision Tree (DT), were selected for comparative analysis. Both models were trained on the 2-dimensional PCA-transformed features.

Hyperparameter Tuning (5-Fold Cross-Validation)

Hyperparameter tuning was conducted using 5-fold cross-validation on the training set, with accuracy as the scoring metric. The results are summarized in Table 1.

Model	Hyperparameter	Mean CV Accuracy	Optimal Setting
Logistic Regression	$C \in \{0.1, 1.0\}$	0.7337 ($C = 0.1$)	$C = 0.1$
Decision Tree	$\text{max\_depth} \in \{3, \text{None}\}$	0.7119 ($\text{max\_depth} = 3$)	$\text{max\_depth} = 3$

Table 1. Cross-Validation Results for Model Selection.

The Logistic Regression model, with a regularization strength of $C = 0.1$, achieved the highest mean cross-validation accuracy.

Test Set Performance and ROC Analysis

The optimally tuned LR ($C = 0.1$) and DT ($\text{max\_depth} = 3$) models were evaluated on the independent hold-out test set. The full performance metrics are presented in Table 2.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.7403	0.7467	0.9825	0.8485	**0.7425**
Decision Tree	0.7013	0.7429	0.9123	0.8189	**0.5259**

Table 2. Comparative Test Set Performance Metrics.

The comparative analysis, supported by the Receiver Operating Characteristic (ROC) curve analysis, clearly indicates the superiority of the Logistic Regression model. The LR model achieved an **Area Under the Curve (AUC) of** 0.743, demonstrating good discriminatory power. In contrast, the Decision Tree model achieved an AUC of 0.526, indicating performance only slightly better than random chance.

4. Final Recommendation

The **Logistic Regression model** is recommended as the final predictive model. Its superior AUC and more balanced performance metrics on the test set demonstrate better generalization capability. Furthermore, the high Recall score (0.982) indicates that the LR model is highly effective at correctly identifying patients who survived, a critical requirement in a medical context where false negatives (predicting death when the patient survived) should be minimized. The Decision Tree's near-random AUC suggests it failed to learn the underlying patterns effectively, likely due to the small, linearly separable nature of the PCA-transformed dataset.

References

Haberman, S. J. (1976). *Haberman's Survival Data*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/43/haberman+s+survival