Mohammad Bagher Soroush - 400130273 - ML 1 Report

1 .Explanation of Process Steps

The analysis process includes data preprocessing, model training, and evaluation. The steps are as follows:

Data Loading and Inspection: The dataset was loaded and inspected to understand its structure, data types, and the presence of any categorical variables.

Data Preprocessing: One-hot encoding was applied to convert categorical columns (fem and mar) into numeric columns suitable for modeling.

Model Training: The dataset was split into training and test sets. An LDA model was then trained to classify individuals based on the processed features.

2 .Why One-Hot Encoding Was Used

One-hot encoding was essential because the dataset contained categorical variables, such as gender (fem) and marital status (mar). These categorical variables cannot be directly used in the LDA model, as it requires numeric input. By applying one-hot encoding, each category within the categorical columns was converted into binary columns (0 or 1), allowing the model to interpret these categories as distinct features.

3 .Model Results Interpretation

The model was evaluated on test data, and various classification metrics were used to interpret its performance:

Accuracy Score: Indicates the percentage of correct predictions out of the total predictions. High accuracy implies effective performance, especially if the dataset is balanced.

Precision and Recall: Precision indicates the accuracy of positive predictions, while recall reflects the model’s ability to capture all true positives. These metrics are crucial in evaluating class-wise performance.

F1 Score: The F1 score is a harmonic mean of precision and recall, useful for understanding the balance between precision and recall, especially if the classes are imbalanced.

4 .ROC Curve Interpretation

The ROC curve was plotted to evaluate the model's discrimination threshold, showing the trade-off between the true positive rate (sensitivity) and the false positive rate. The closer the ROC curve is to the top left corner, the better the model's performance. The Area Under the Curve (AUC) provides a single metric to assess the overall effectiveness, where values closer to 1 indicate strong model discrimination.