"Heart Failure Risk Prediction: A Machine Learning Approach"

Mansi yadav 2445828

WHAT IS IT?

Heart disease encompasses a range of conditions affecting the heart, posing significant health challenges globally. Early detection is critical in mitigating severe outcomes and optimizing treatment plans. With advancements in machine learning, predictive models offer new avenues to aid healthcare professionals.

This project utilizes logistic regression, a robust and interpretable machine learning technique, to predict heart disease, employing dimensionality reduction and clustering techniques to enhance understanding of the dataset..

PROBLEM STATEMENT

Heart disease remains a leading cause of mortality worldwide. There is a need for reliable, non-invasive, and data-driven methods to predict and manage heart disease risk effectively. This project addresses this gap by leveraging machine learning techniques to develop a predictive model for early intervention and improved healthcare outcomes.

RESULT

Results (Logistic Regression):

1. Accuracy: 83.33%
2. Best Dimensionality Reduction:
   - LDA achieves the highest accuracy (84%), outperforming PCA (76%) and ICA (76%).
3. Metrics:
   - Class 0 (Non-PCOS): Precision = 75%, Recall = 89%, F1-Score = 81%
   - Class 1 (PCOS): Precision = 92%, Recall = 79%, F1-Score = 85%
   - Macro Avg: Precision = 83%, Recall = 84%, F1-Score = 83%
4. ROC-AUC Score: 0.91 .

METHODOLOGY

DATA PREPROCESSING

- Managed missing data and normalized numerical features like age, cholesterol, and blood pressure.
- Separated numerical and categorical columns for proper transformations.
- Encoded categorical variables such as chest pain type and exercise-induced angina using label encoding.

Model Development:

- Built a Logistic Regression model, which uses a sigmoid function to predict probabilities for binary classification (PCOS or non-PCOS).
- Tuned hyperparameters to optimize performance and prevent overfitting.

CORRELATION MATRIX

Examined relationships between features (e.g., cholesterol vs. heart disease) to identify key predictors and reduce multicollinearity.

DIMENSIONALITY REDUCTION:

Principal Component Analysis (PCA):

- Standardized the dataset to ensure all features contribute equally to variance
- Identified the principal components that explain 95% of the variance using cumulative explained variance.
- Reduced the dataset to the selected components, simplifying analysis and improving model performance.

Linear Discriminant Analysis (LDA):

- Used to project data onto a lower-dimensional space while maximizing the separability between predefined classes.
- LDA works by finding a linear combination of features that minimizes within-class variance and maximizes between-class variance, making it effective for supervised classification tasks

Independent Component Analysis (ICA) :

- can be used to extract independent features from complex, high-dimensional medical data (e.g., ECG signals, genetic data, or clinical records).
- By identifying independent components, ICA helps to reduce noise and improve the performance of predictive models, leading to more accurate identification of patterns that may be indicative of heart disease.

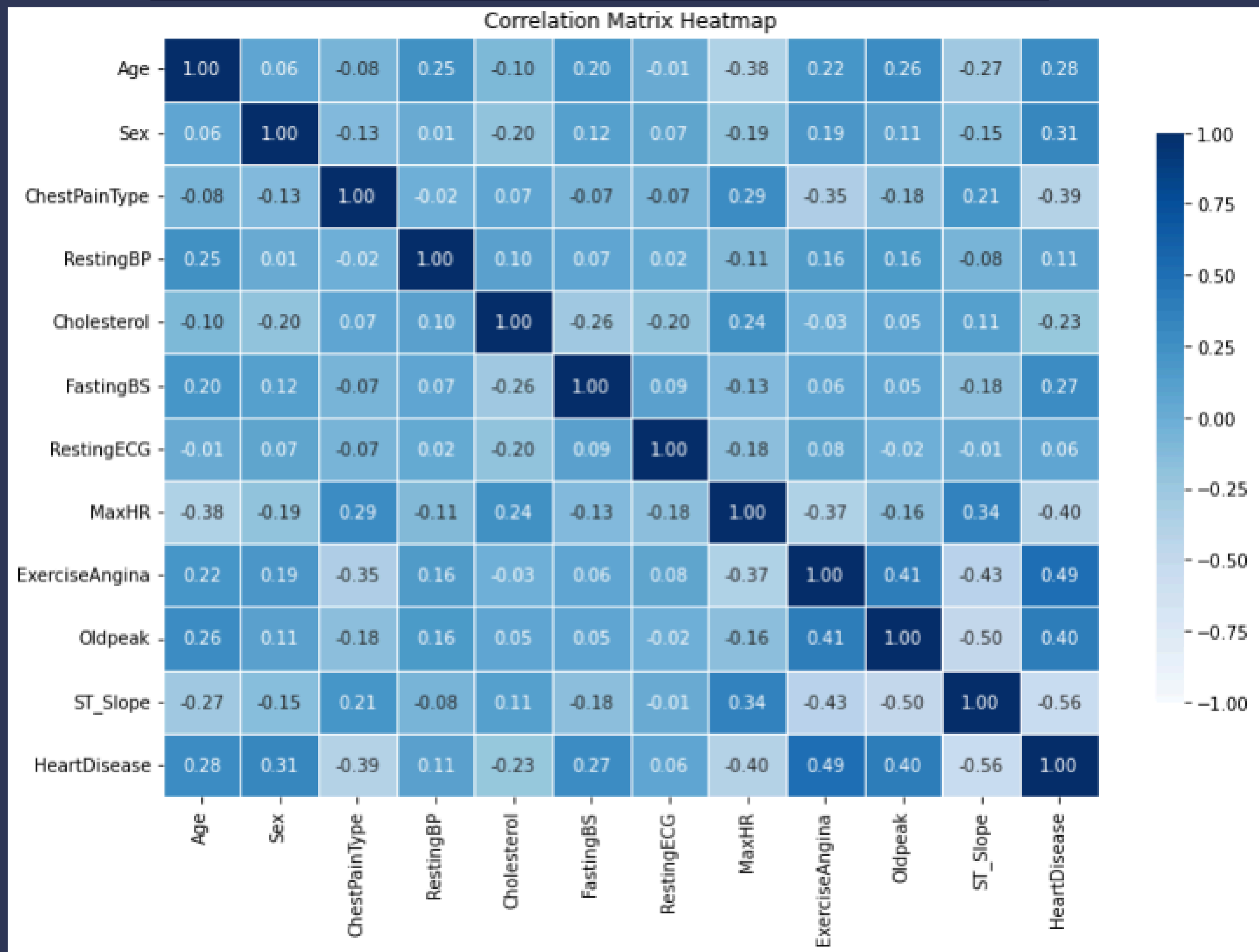
CONCLUSION

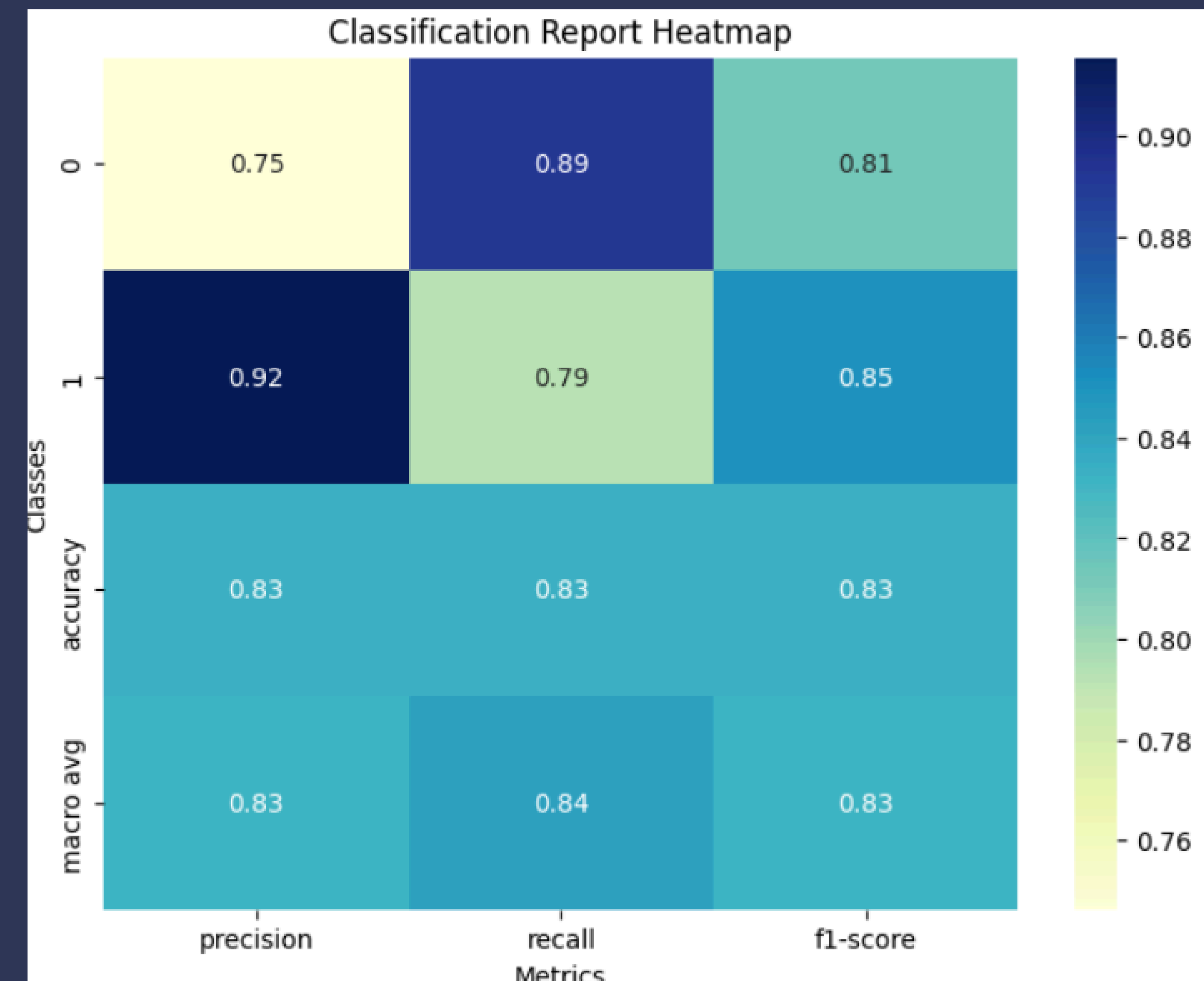
The developed logistic regression model demonstrates strong potential in accurately detecting PCOS based on ultrasound data.

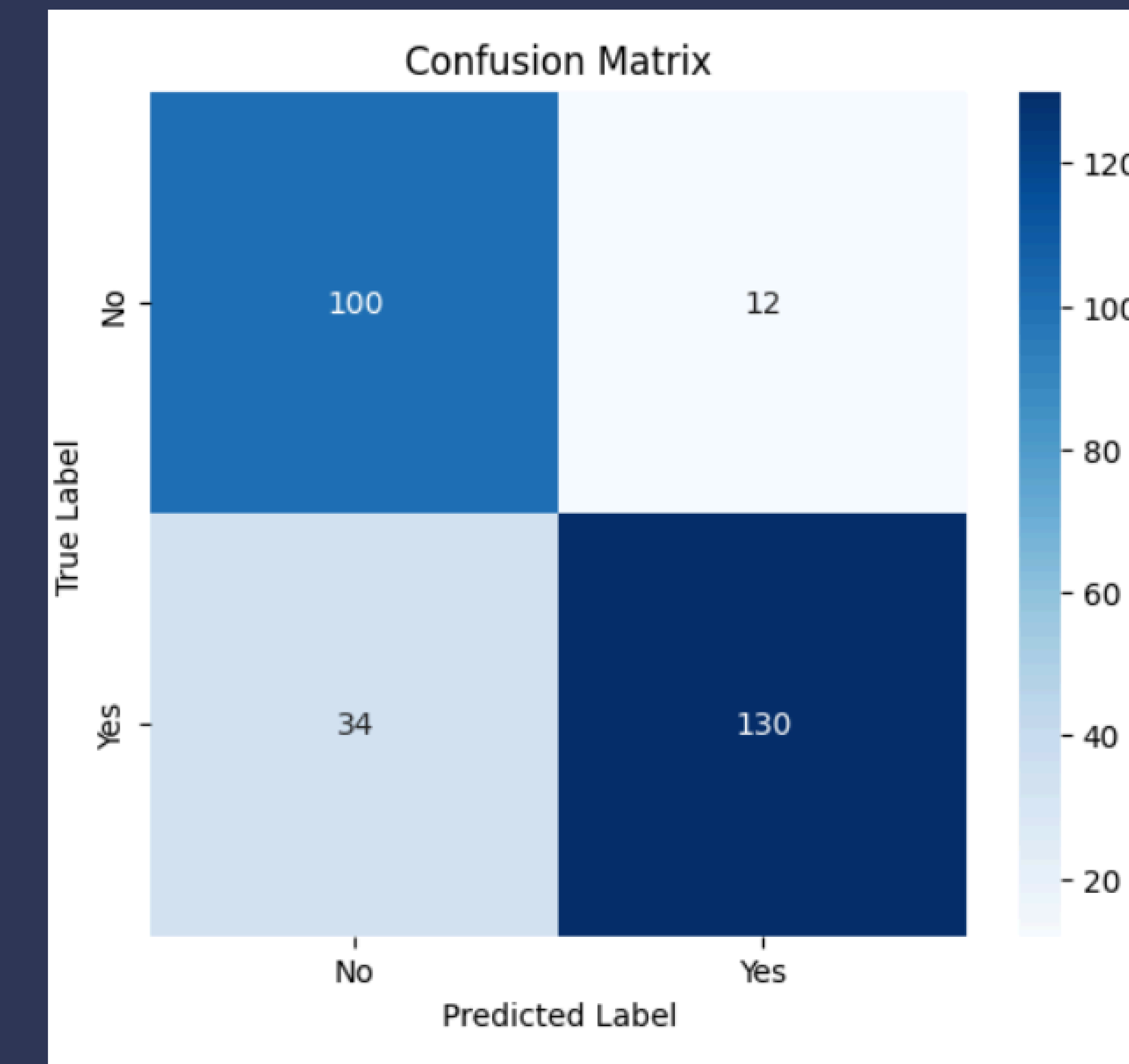
The model achieved an accuracy of 83.33% and an excellent ROC-AUC score of 0.91, indicating reliable classification performance.

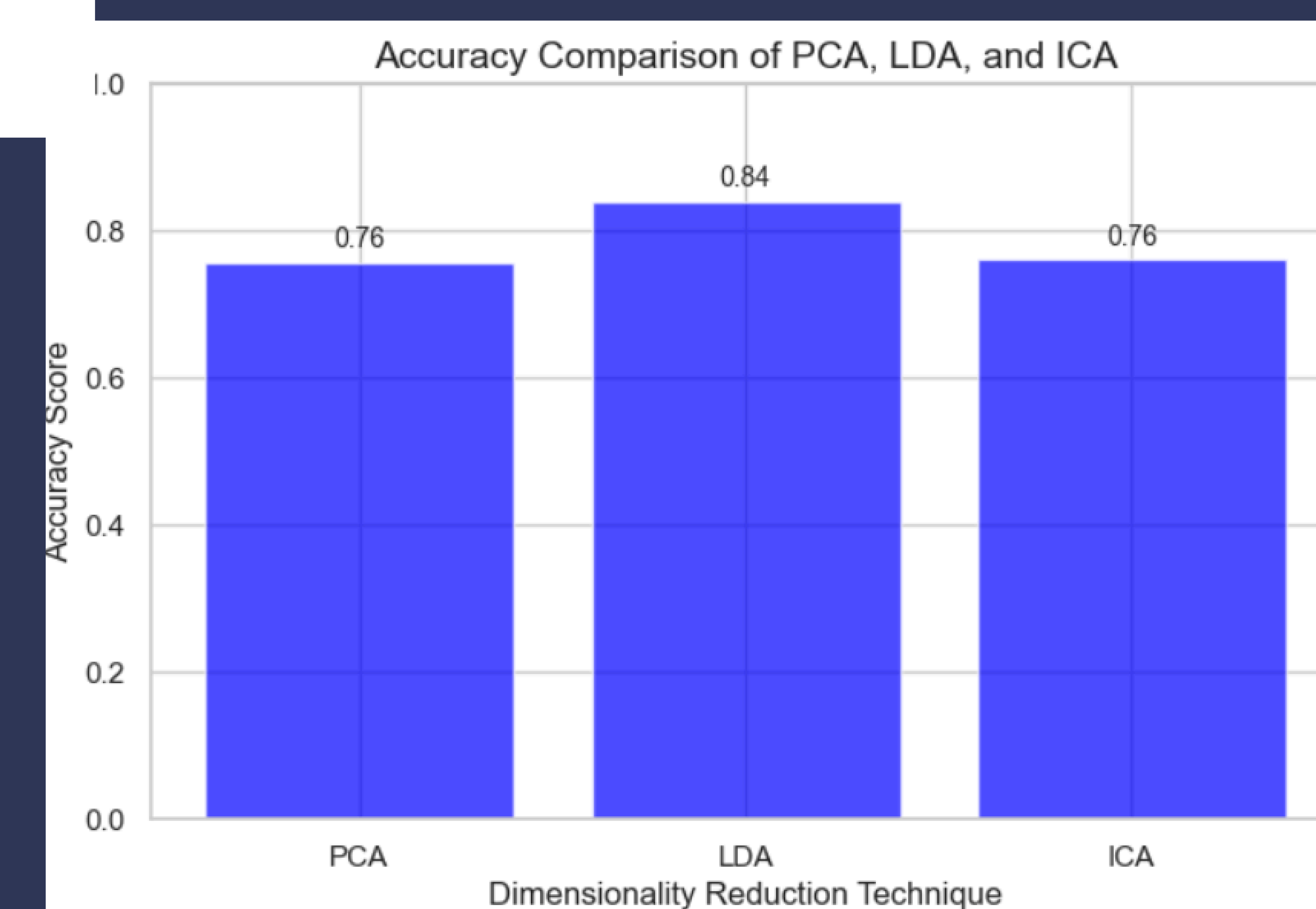
Among dimensionality reduction techniques, LDA performed the best, achieving 84% accuracy, highlighting its effectiveness in preserving class separability.

CORRELATION MATRIX









DISTRIBUTION ACROSS TARGET VARIABLE

CHARTS