PREDICTING HEART DISEASE USING MACHINE LEARNING

Rupesh Rangwani
Dharmik Patel
Aryan Jain
Sanskar Srivastava
Emad Bhaktari
Cagri Isilak

AGENDA

01
Introduction and Objectives

02
Data Exploration

03
Data Preprocessing and Feature Engineering

04
Model Implementation

05
Model Evaluation and Results

06
Insights, Recommendations, and Future Work

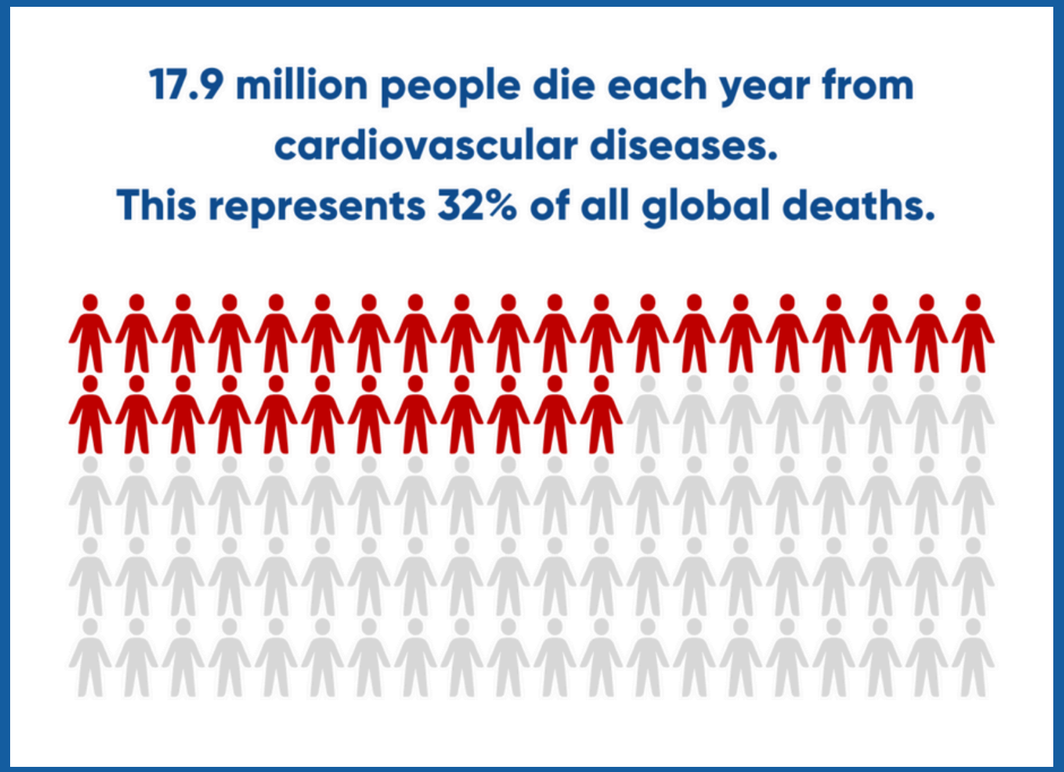
INTRODUCTION & OBJECTIVES

Context

- Cardiovascular diseases account for **32%** of annual deaths globally.
- Early detection is critical for reducing mortality.

Objectives

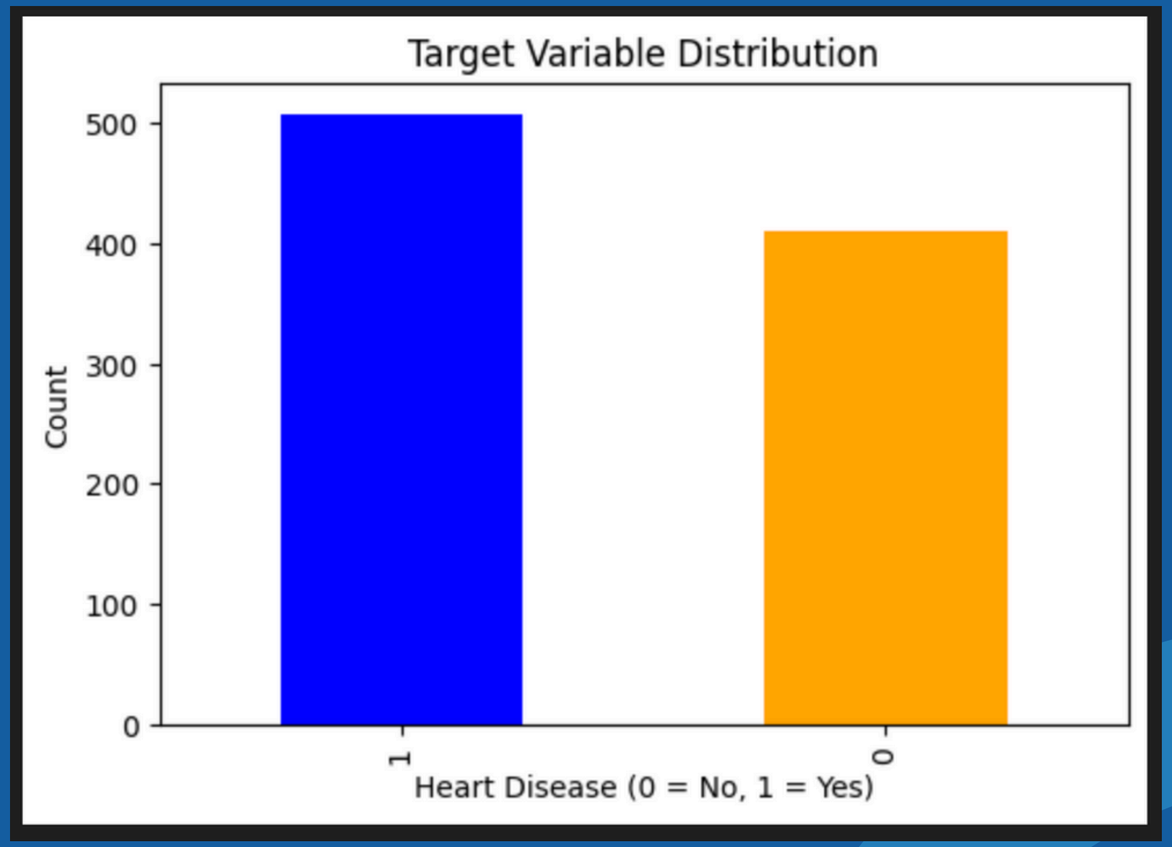
- Explore and preprocess dataset.
- Engineer features for better model performance.
- Evaluate and compare machine learning models.



DATA EXPLORATION

Key Highlights

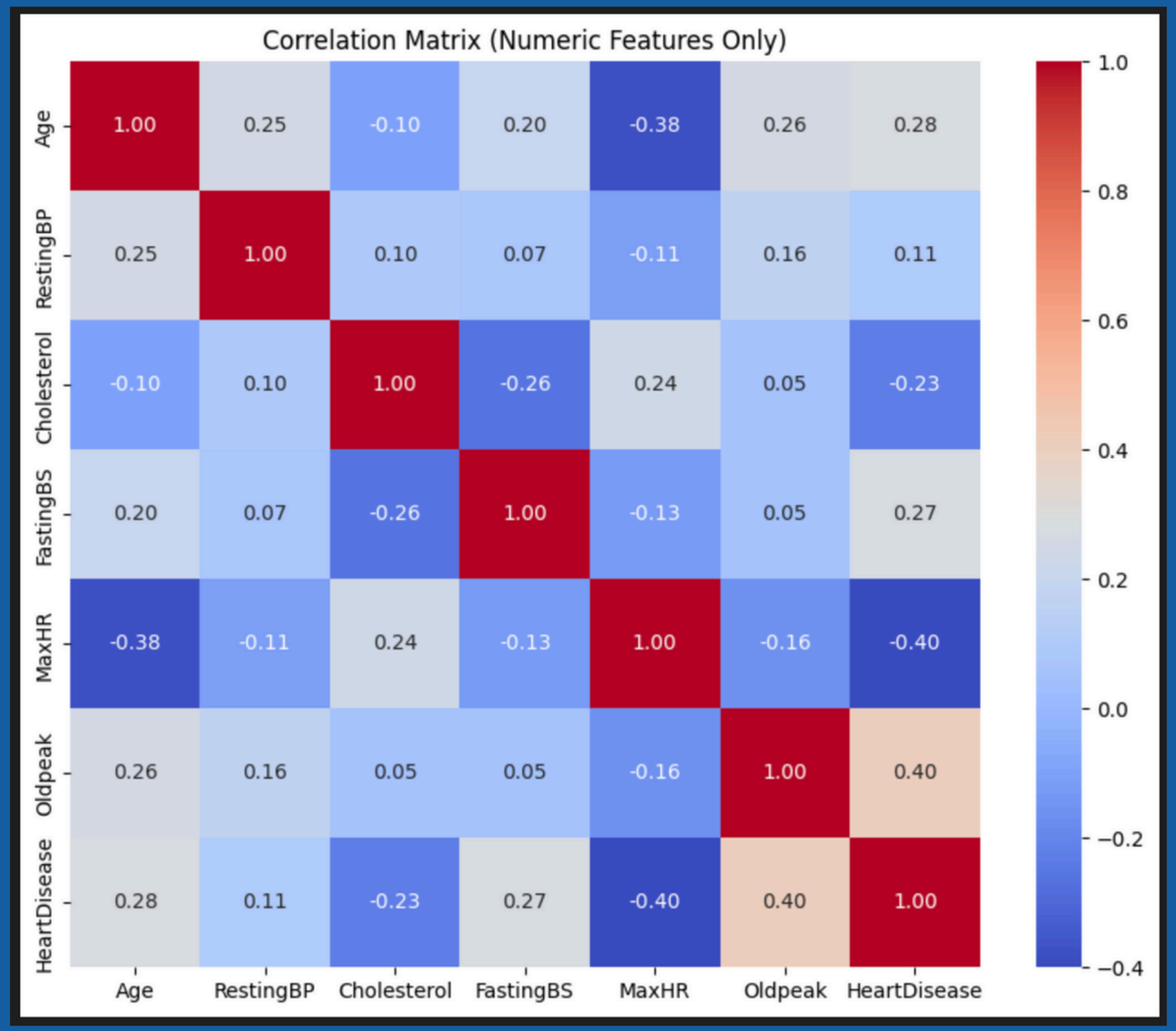
- Dataset from Kaggle with 989 rows & 12 features.
- Binary target variable (HeartDisease: 0 or 1).
- Observed slight class imbalance and correlations between features like ST_Slope and MaxHR.



CORRELATIONS IN DATA

Insights

- High correlation between Oldpeak and Heart Disease.
- MaxHR negatively correlated with Heart Disease.
- Moderate correlation between Age and Heart Disease.



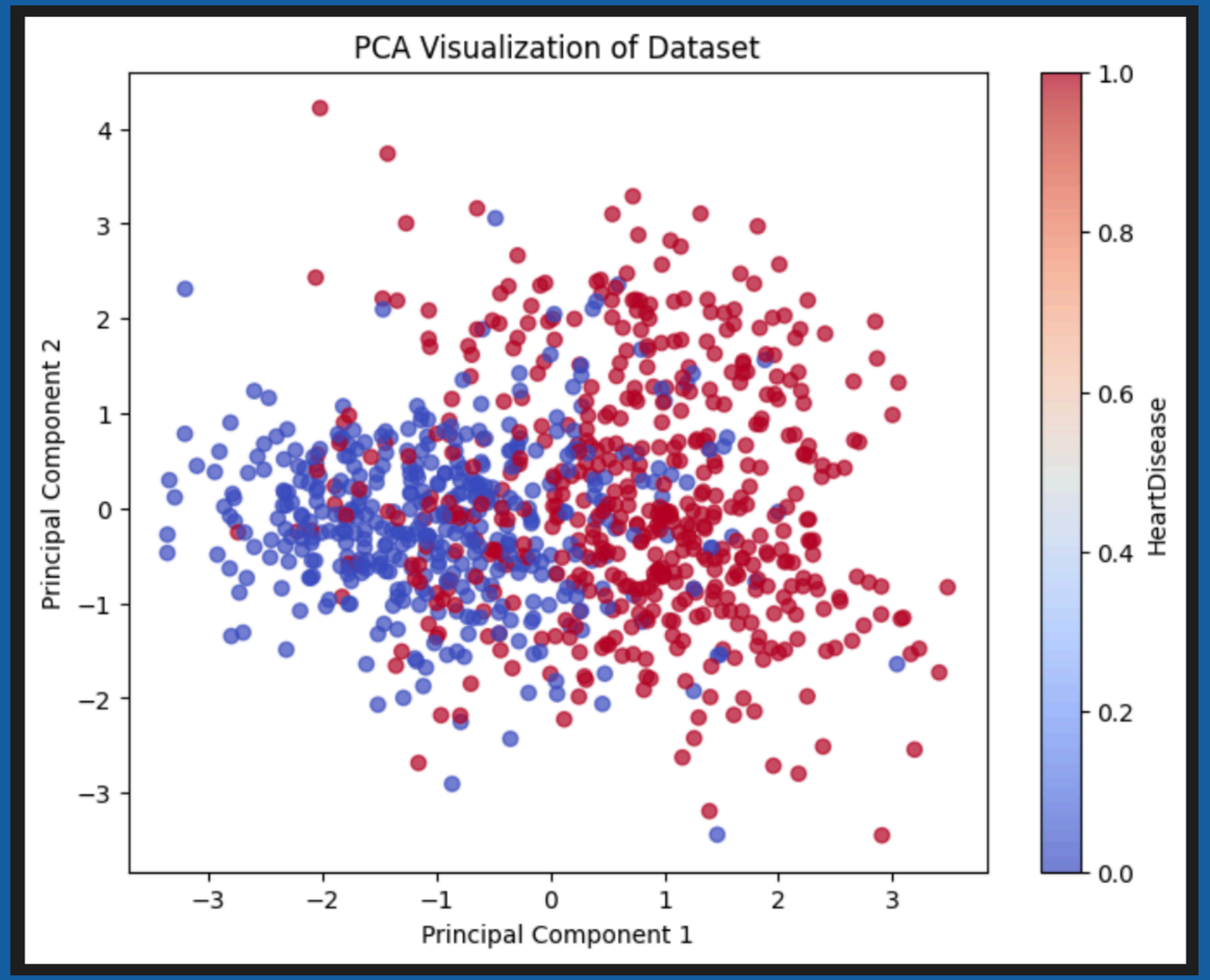
PCA VISUALIZATION

Purpose

- Visualize dataset in reduced dimensions to understand separability of classes.

Insights

- PCA shows moderate class separability in two components.

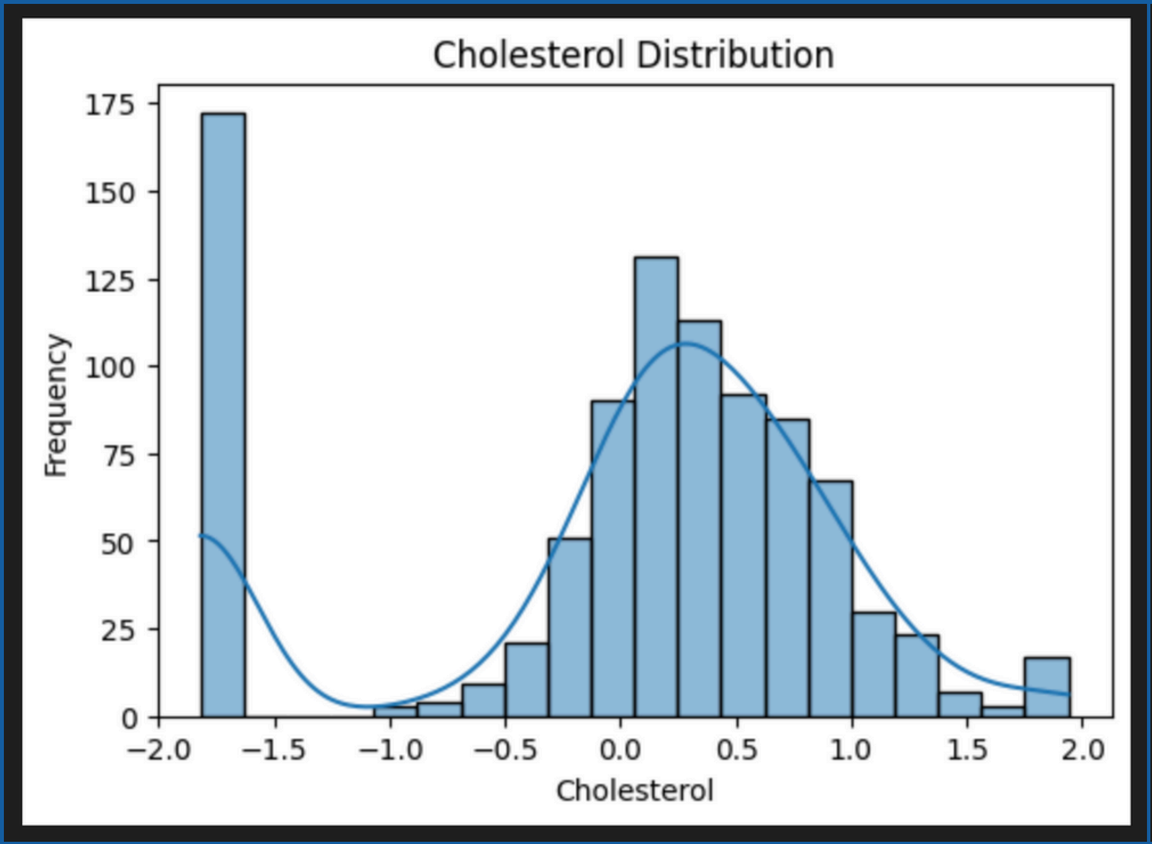


DATA PREPROCESSING

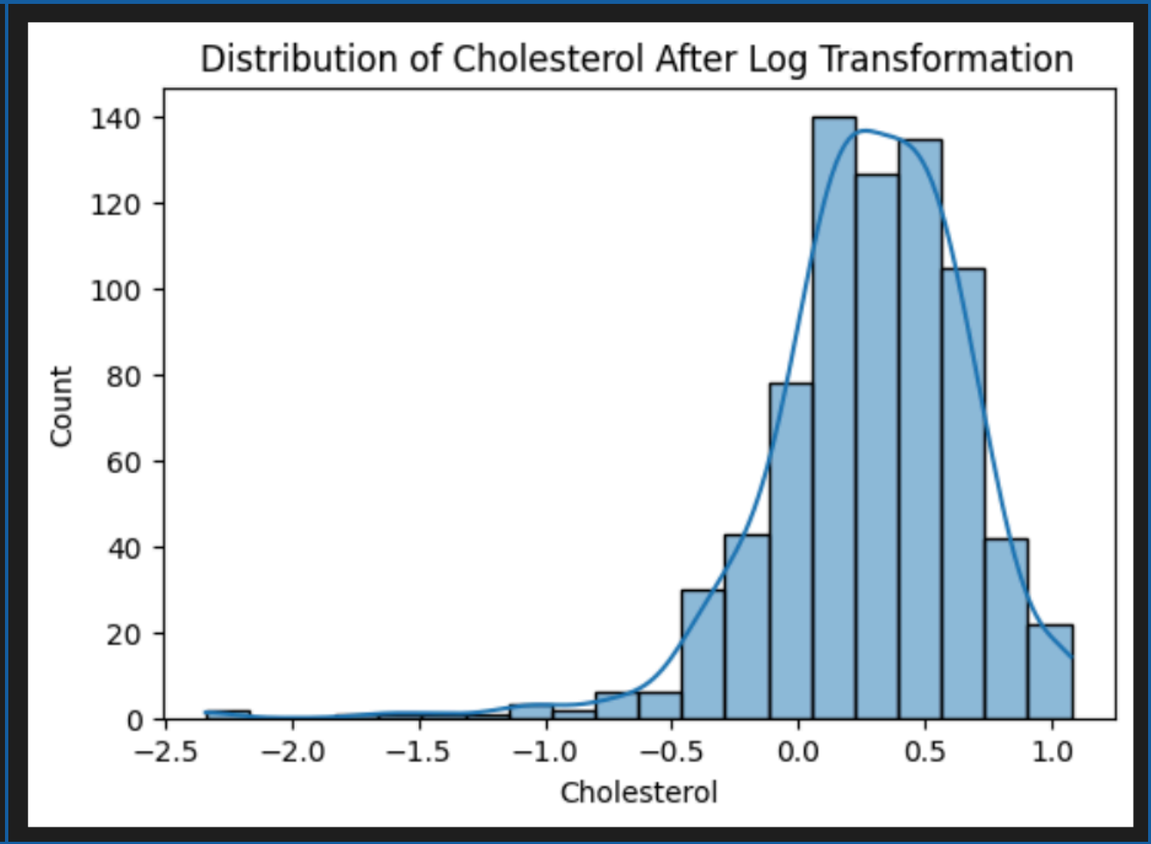
Step Taken

- Handled outliers using percentile capping.
- Applied log transformations to skewed features (e.g., Cholesterol, Oldpeak).
- Standardized numeric features using StandardScaler.

BEFORE



AFTER



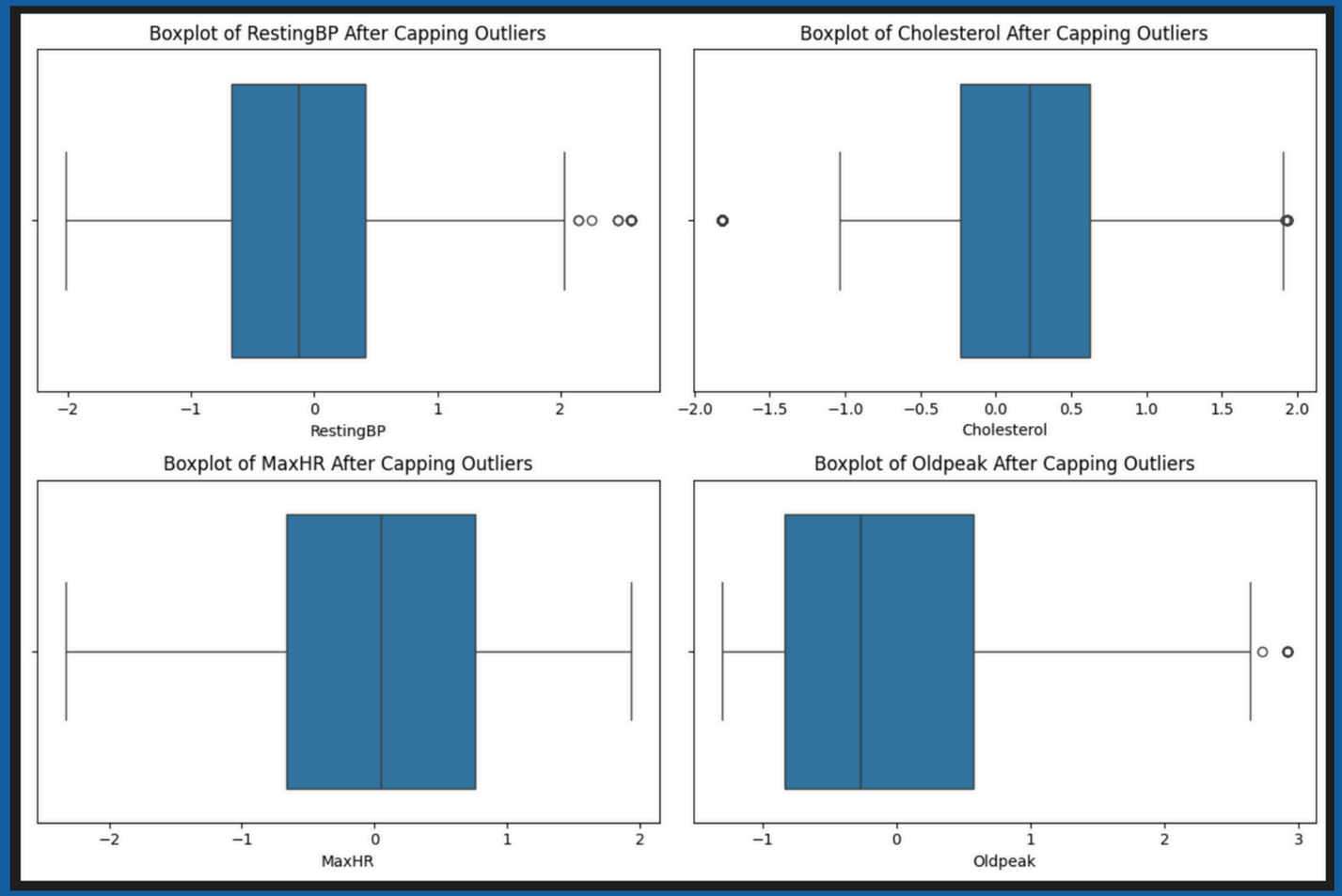
OUTLIER HANDLING

Details

- Outliers in features like RestingBP and Cholesterol capped to 1st and 99th percentiles.

Outcome

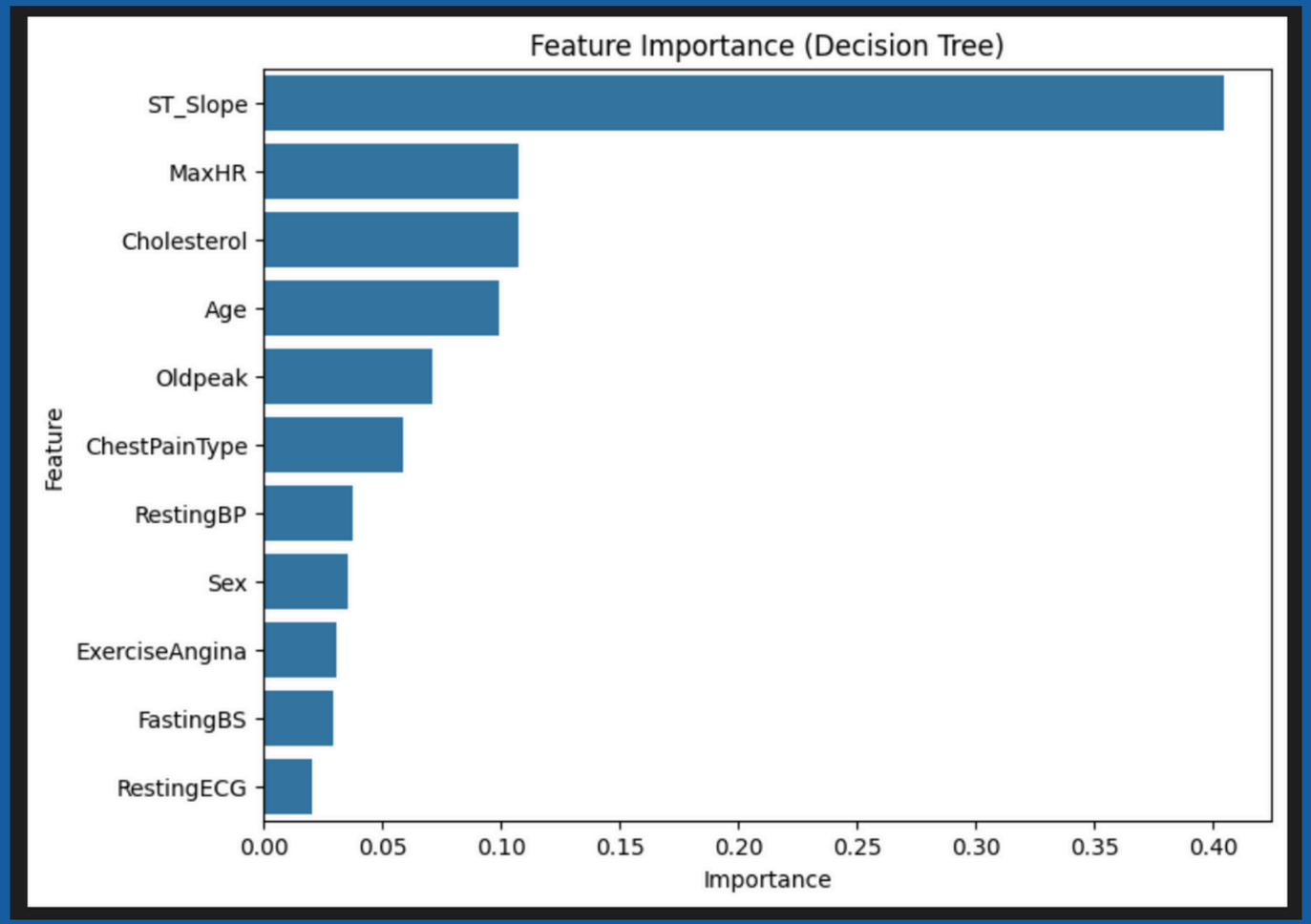
- Improved feature distributions and reduced noise.



FEATURE IMPORTANCE

Top Predictors

- ST_Slope, MaxHR, and Cholesterol were identified as key predictors.
- Feature importance is calculated using Decision Tree.



MODEL IMPLEMENTATION

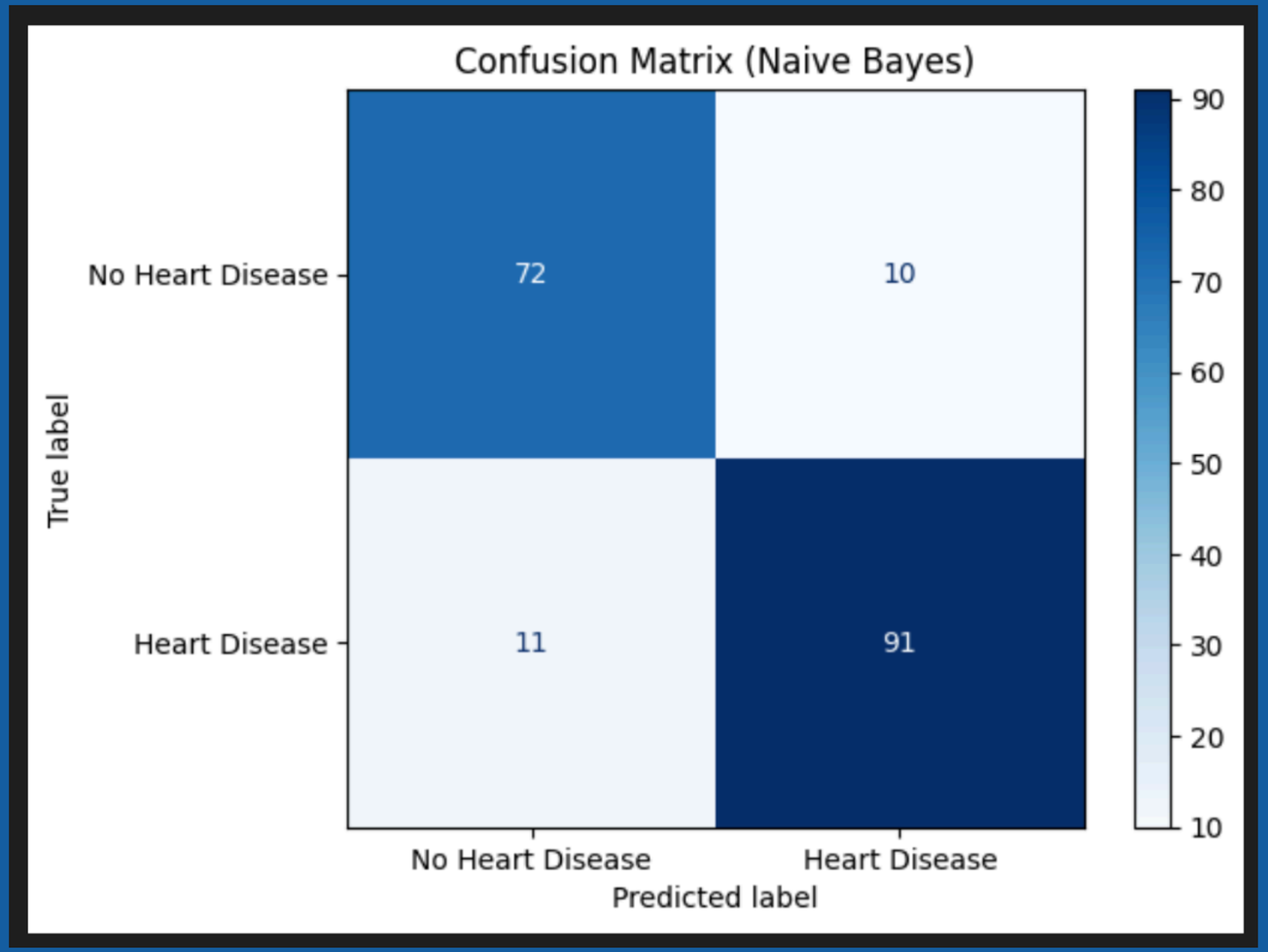
NAIVE BAYES

Model Details

Gaussian Naive Bayes with var_smoothing = 1e-9.

Performance

- Accuracy: 89%
- F1-Score: 89%
- AUC: 0.94 (highest).



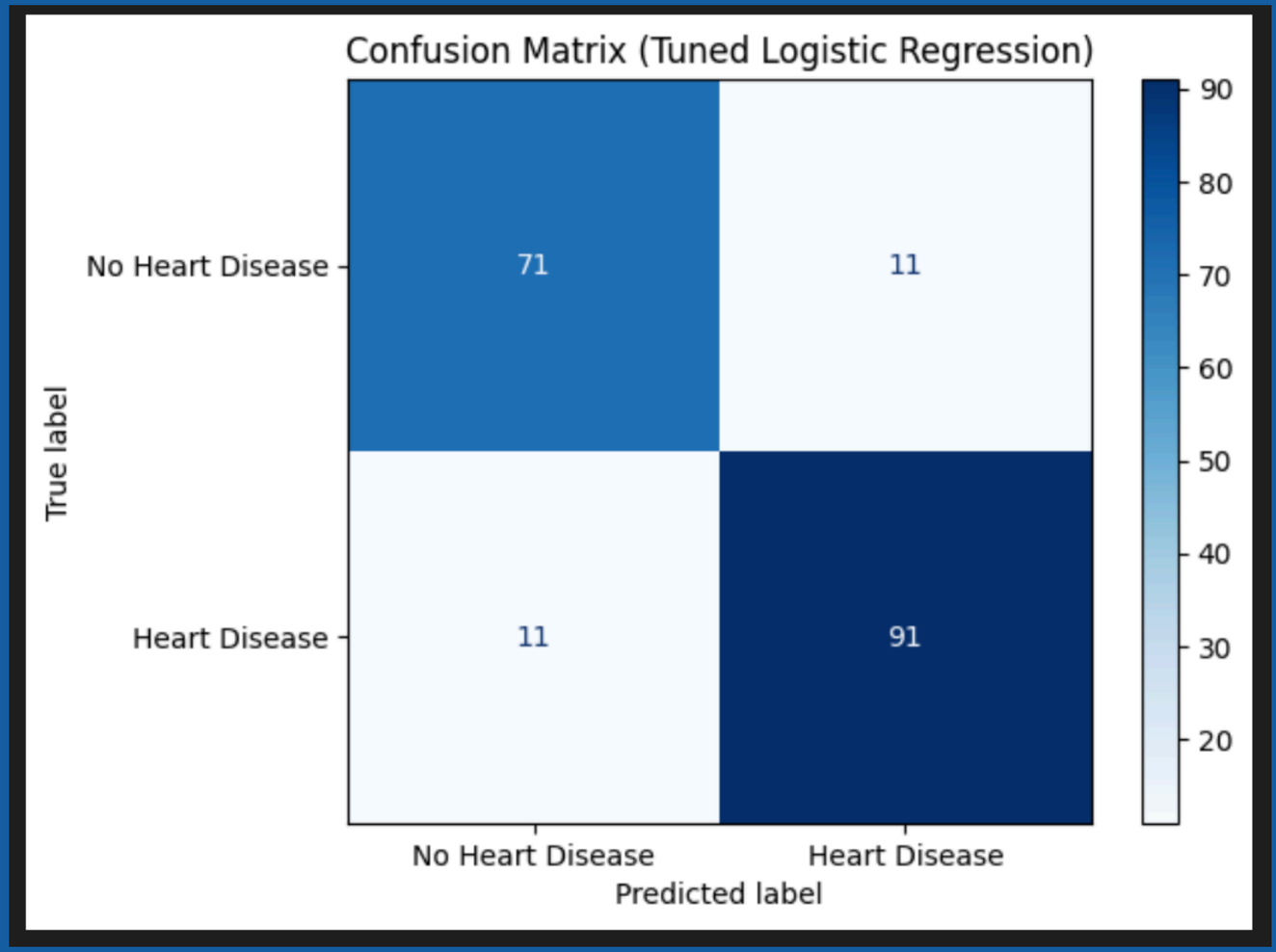
LOGISTIC REGRESSION

Model Details

Logistic Regression with L2 regularization.

Performance

- Accuracy: 88%
- Precision: 0.91
- Recall: 0.84
- F1-Score: 0.88
- AUC: 0.91



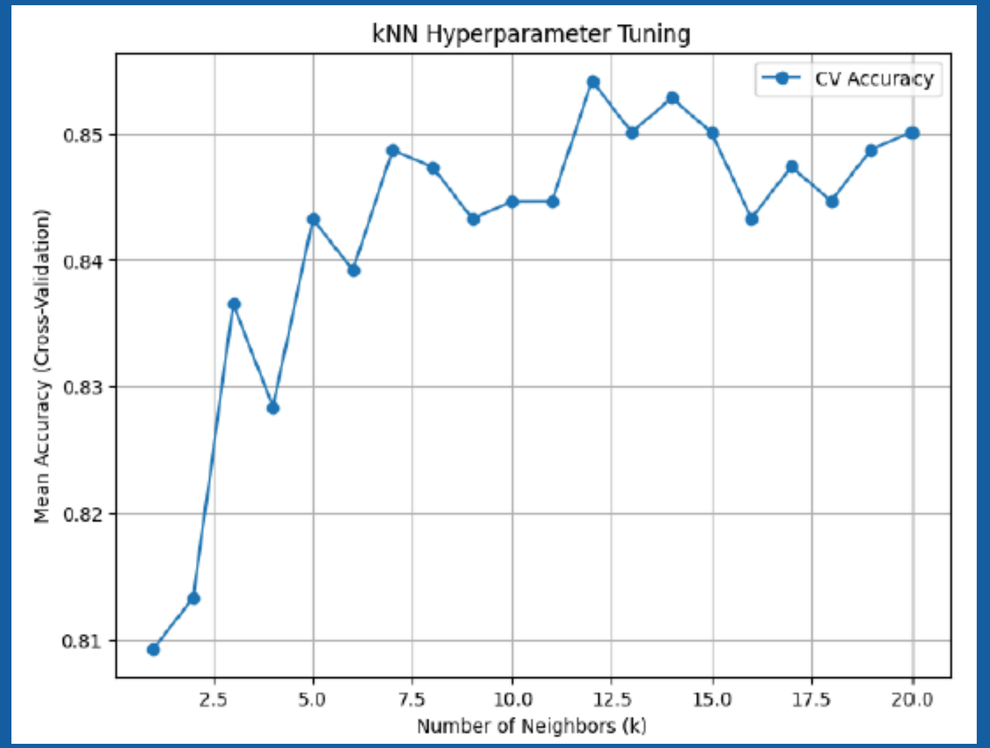
K-NEAREST NEIGHBORS (KNN)

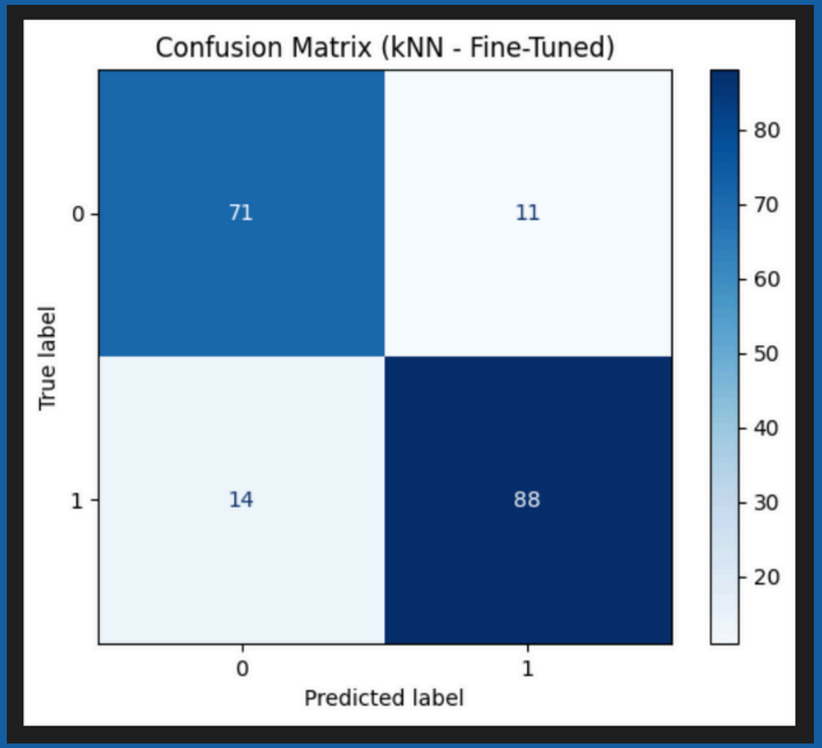
Model Details

Optimal k = 12 (GridSearchCV).

Performance

- Accuracy: 87%
- Precision: 0.88
- Recall: 0.89
- F1-Score: 0.88
- AUC: 0.92





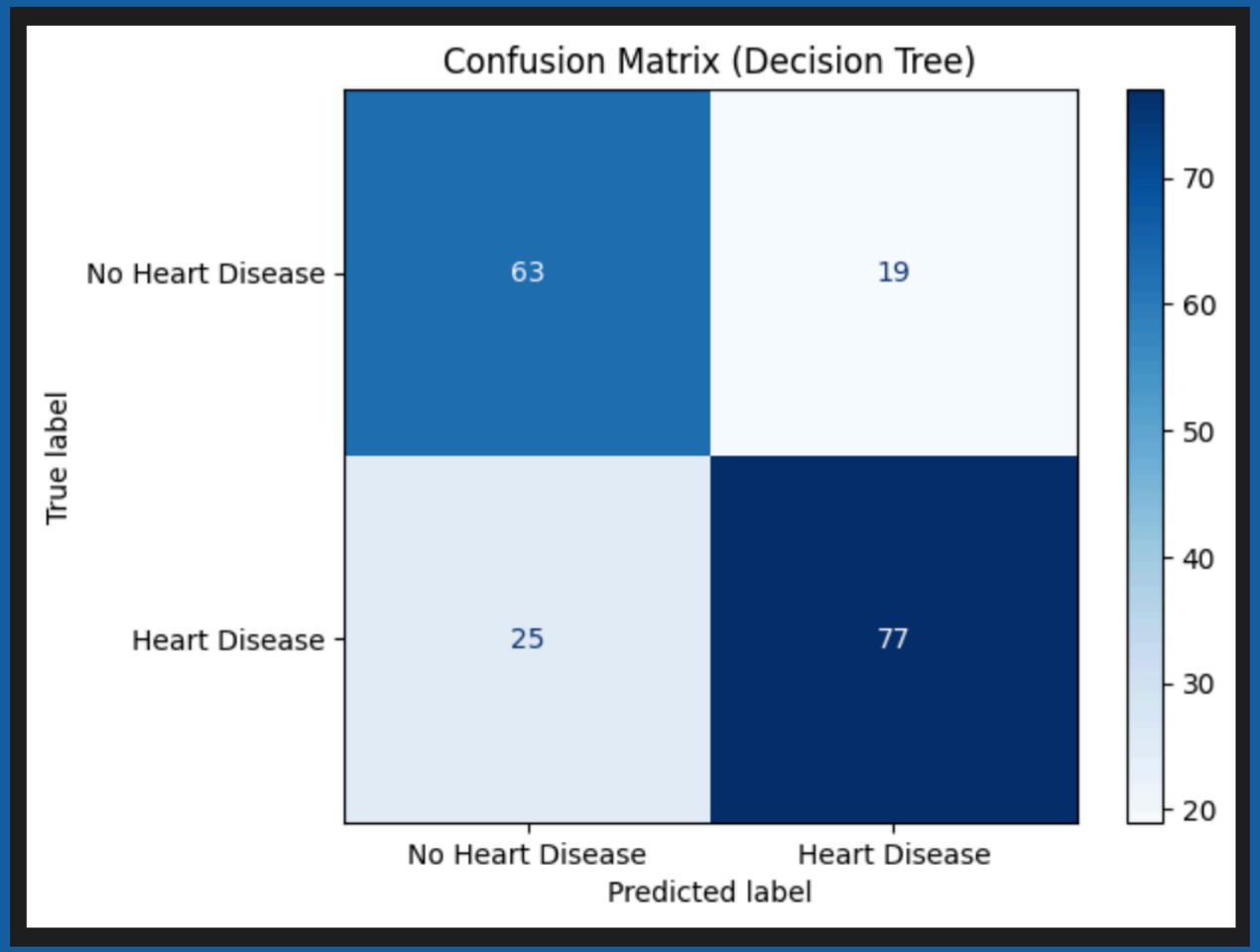
DECISION TREE

Model Details

Max depth = 3 (optimized).

Performance

- Accuracy: 79%
- Precision: 0.78
- Recall: 0.77
- F1-Score: 0.79
- AUC: 0.84



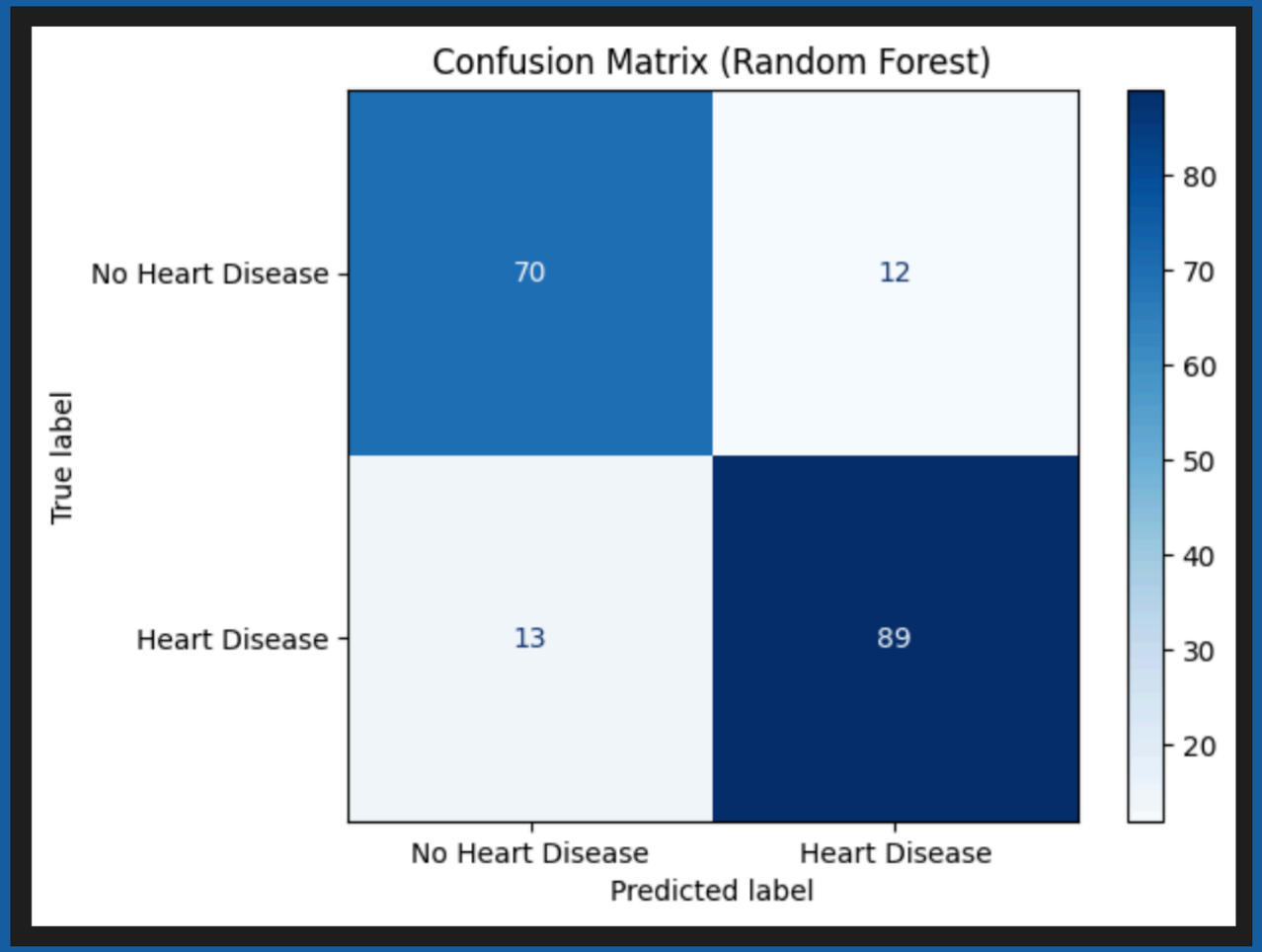
RANDOM FOREST

Model Details

Max depth = 10, n_estimators = 50.

Performance

- Accuracy: 87%
- Precision: 0.91
- Recall: 0.86
- F1-Score: 0.88
- AUC: 0.92



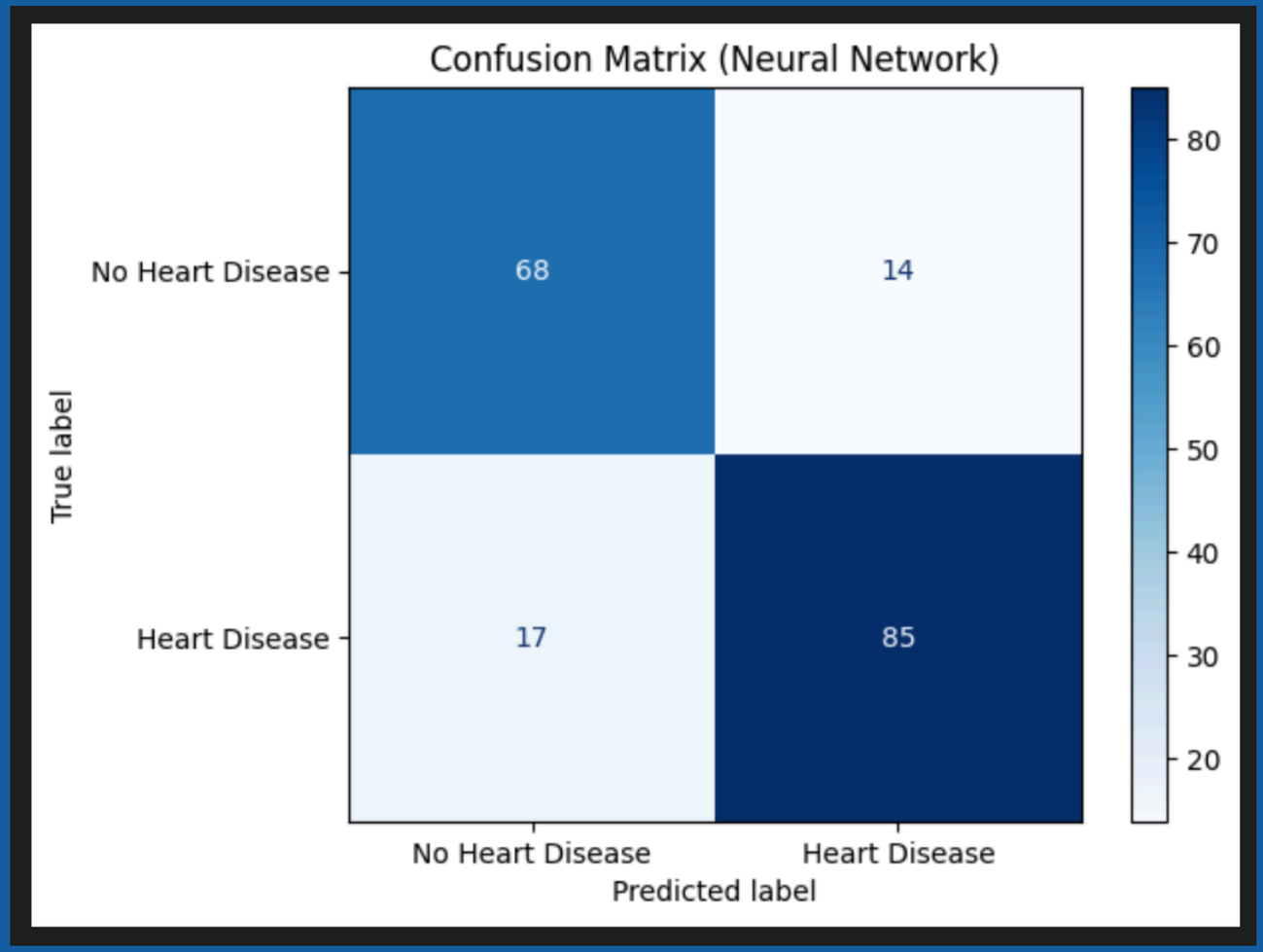
NEURAL NETWORK

Model Details

Architecture: Three-layer (64-32-1), ReLU and Sigmoid activations.

Performance

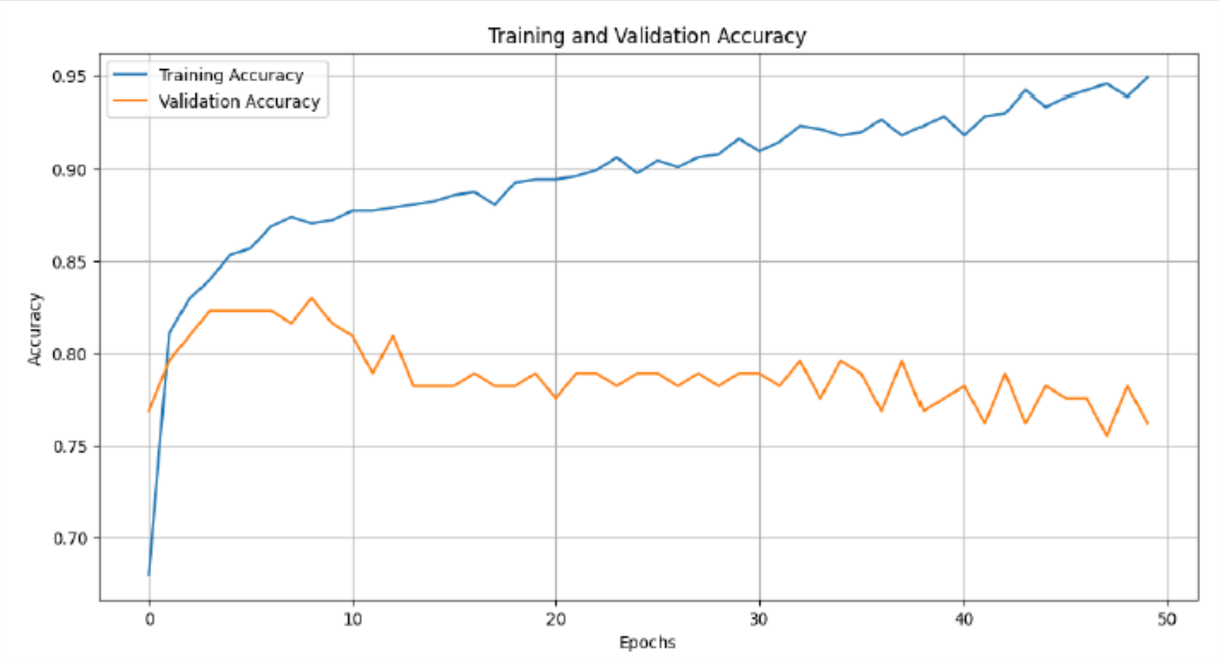
- Accuracy: 88%
- Precision: 0.95
- Recall: 0.84
- F1-Score: 0.89
- AUC: 0.91



NEURAL NETWORK

Training Observations

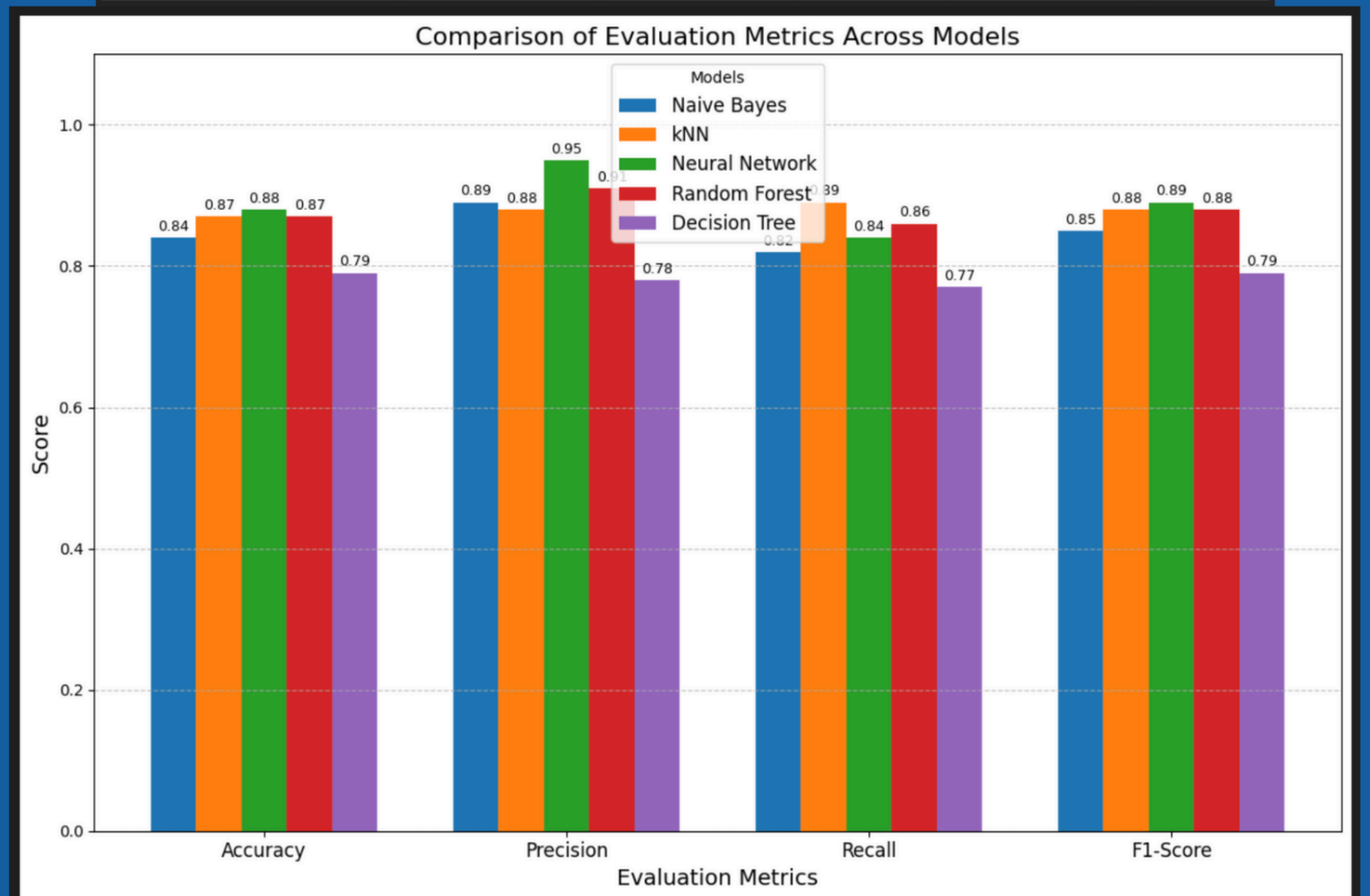
Validation accuracy lags slightly behind training accuracy, indicating potential overfitting.



COMPARISON OF MODELS

COMPARISON OF MODELS

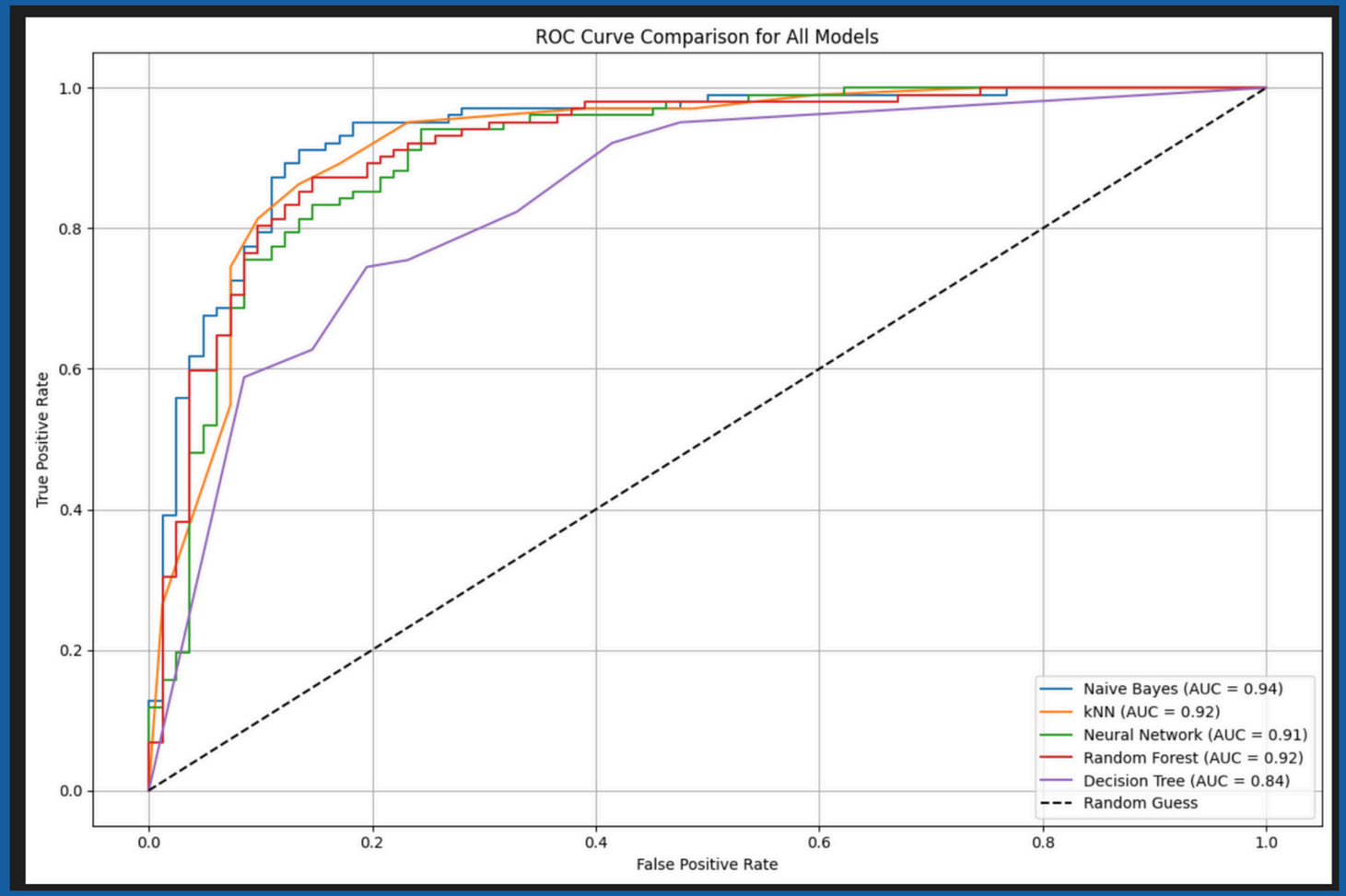
Model	Accuracy	Precision	Recall	F1-Score	AUC
Naive Bayes	0.84	0.89	0.82	0.85	0.94
kNN	0.87	0.88	0.89	0.88	0.92
Neural Network	0.88	0.95	0.84	0.89	0.91
Random Forest	0.87	0.91	0.86	0.88	0.92
Decision Tree	0.79	0.78	0.77	0.79	0.84



KEY INSIGHTS AND BEST MODEL

Key Insights

- **Naive Bayes**: Highest AUC (0.94), simple but assumes feature independence.
- **Neural Network**: Best Precision (0.95), slightly overfits validation data.
- **Random Forest**: Balanced metrics (Precision: 0.91, Recall: 0.86), ideal for clinical use.
- **Decision Tree**: Underperforms across all metrics.



Best Model

- **Random Forest** for its balance between precision, recall, and ease of interpretability.

RECOMMENDATIONS

Applications

- Random Forest for high-recall scenarios in clinical practice.
- Neural Networks for high-accuracy predictions where computational resources are available.
- Naive Bayes for low-resource environments.

FUTURE DIRECTIONS

Improve Dataset

Use larger datasets and address class imbalance.

Explore Models

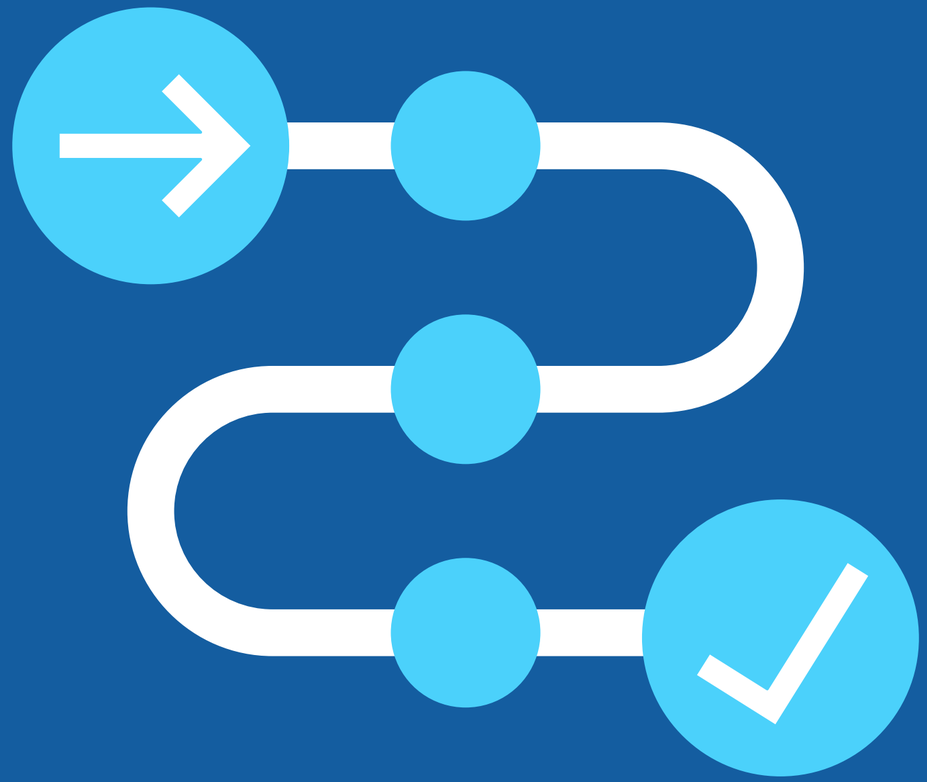
Test advanced methods like XGBoost.

Focus on Explainability

Use SHAP for interpretable results.

Real-World Validation

Deploy models in clinical settings.



Thank You!