Predictive Modeling for Heart Disease Diagnosis Using Machine Learning

- *Authors: Ramiz, Jalil, Parviz, Rashad, Xanahmad*

- *Institution: Khazar University, Computer Science*

- *Date: 05/28/2025*

Introduction

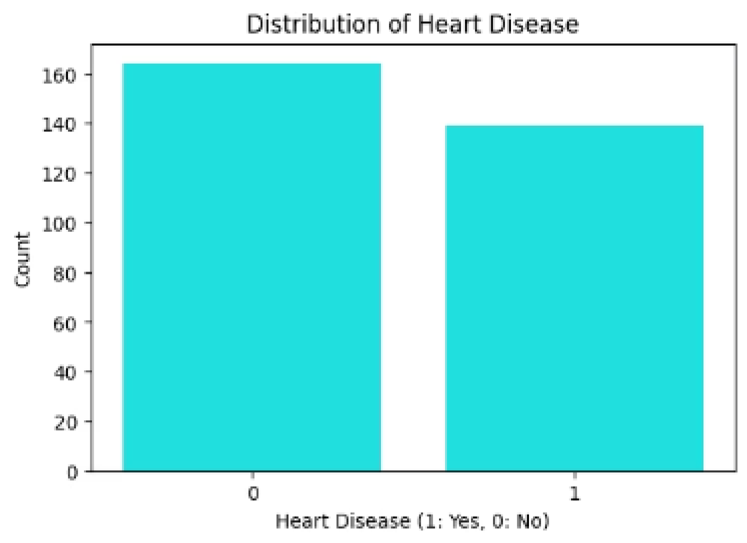
Key Points

- Cardiovascular diseases are the leading cause of global mortality.
- Early diagnosis is crucial for effective treatment.
- Machine learning can assist clinicians in identifying high-risk patients.
- Objective: Build and evaluate ML models for binary classification of heart disease.

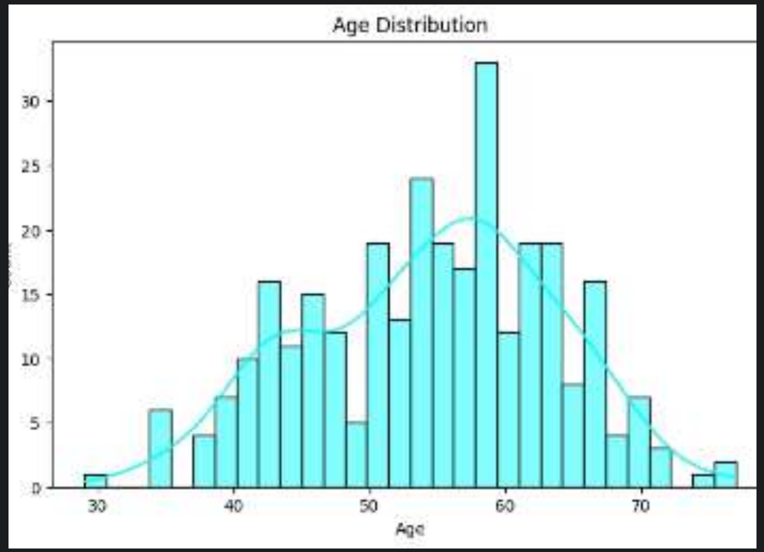
Dataset & Preprocessing

- UCI Heart Disease dataset used (303 instances, 14 features).
- Target: Presence (1) or absence (0) of heart disease.
- Missing values in 'ca' and 'thal' filled using mode.
- Features scaled using StandardScaler.

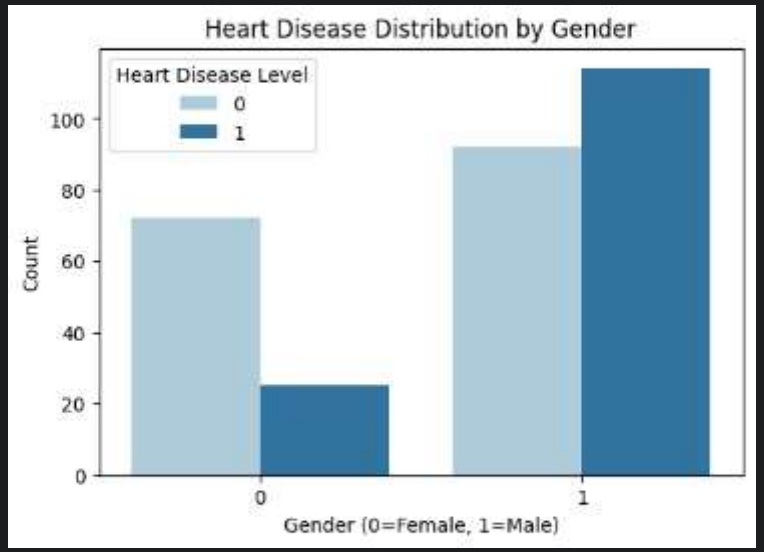
Visualizations



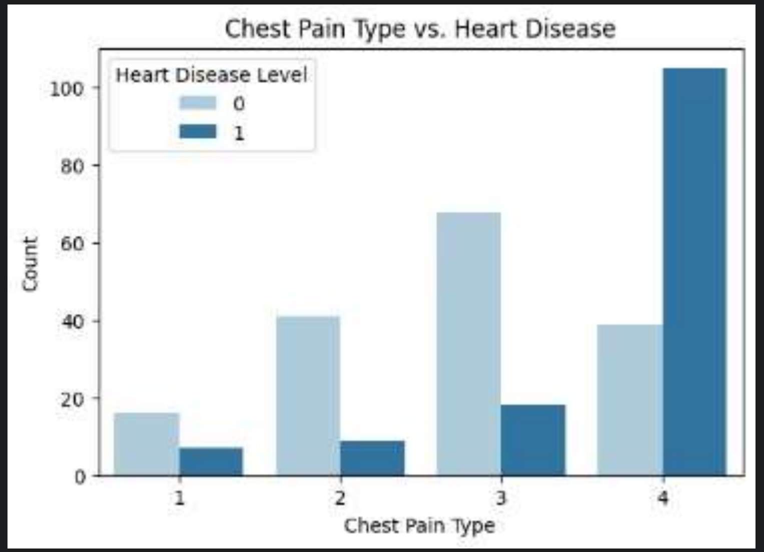
Distribution of classes



Age distribution

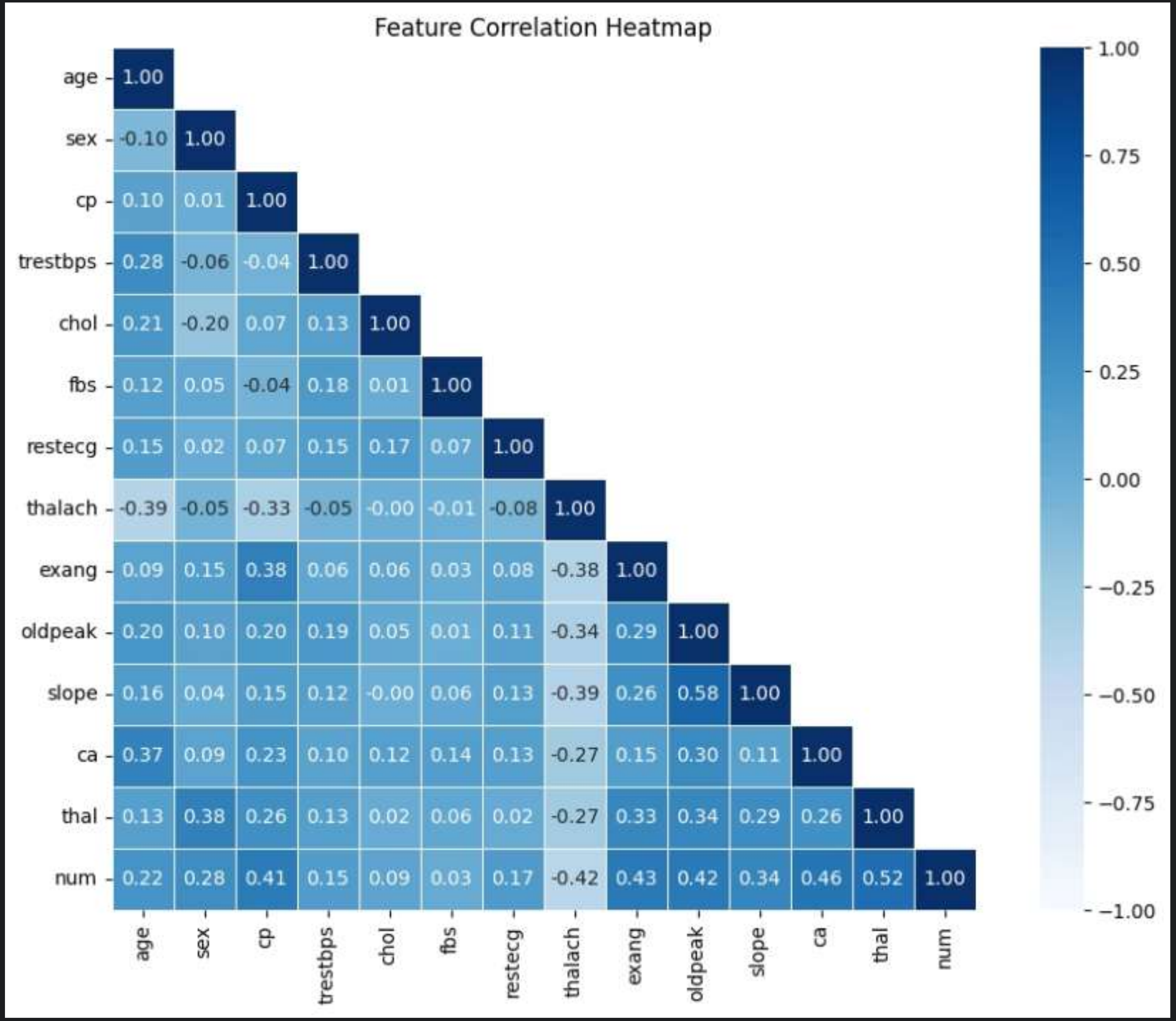


Disease distribution by gender

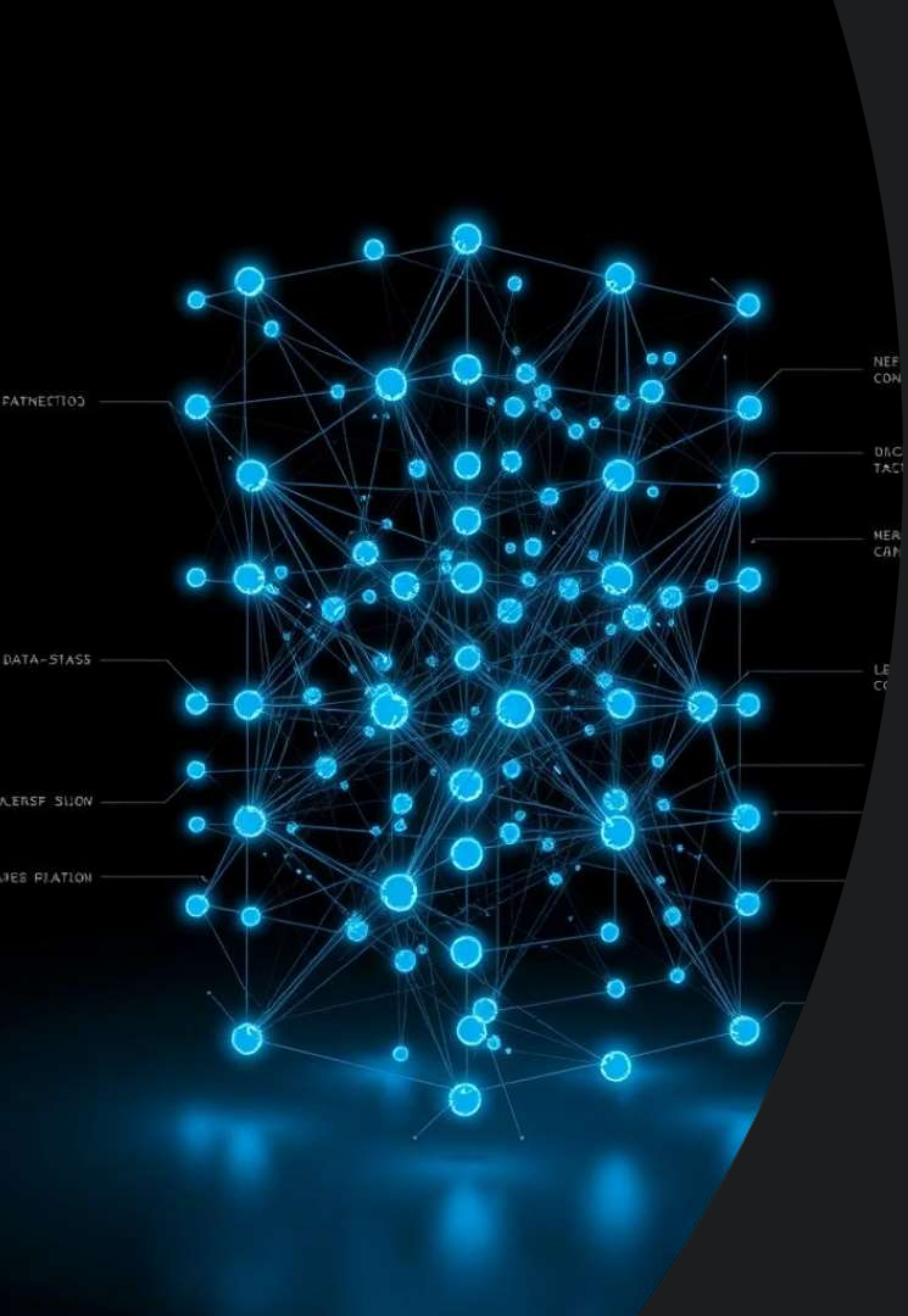


Heart disease by chest pain type

Correlation Heatmap of Heart Disease Features



- **This Feature Correlation Heatmap shows relationships between various health metrics:**
- Strong positive correlations appear between 'num' and several features ('thal' : 0.52, 'ca' : 0.46)



Methodology

(Part 1)

- Models used: Logistic Regression, Random Forest, SVC, KNN.

- Stratified K-Fold Cross-Validation (k=5) applied for balanced evaluation.

- GridSearchCV used for hyperparameter tuning.

- ROC AUC, Precision, Recall used as a key metric.

Methodology

(Part 2)

- Evaluation Metrics: Accuracy, Precision, Recall, F1-score, ROC AUC.

- ROC AUC emphasized due to clinical importance of minimizing false negatives.

- Confusion matrices used to interpret model behavior.

Results Overview

Model	ROC AUC	Precision	Recall	F1-Score	Accuracy
KNN	0.94	0.94	0.91	0.92	0.918
Random Forest	0.94	0.96	0.84	0.90	0.902
SVC	0.93	0.93	0.88	0.90	0.902
Logistic Regression	0.94	0.88	0.91	0.89	0.885

1.Logistic Regression : High recall, interpretable.

1.Random Forest : High precision, robust to overfitting.

1.SVC : Balanced and effective with proper scaling.

1.KNN : Highest accuracy, sensitive to feature scaling.

Real-World Application

- Logistic Regression in clinics needing transparency.
- KNN or SVC where slight accuracy gains are critical.
- Random Forest in automated pipelines with large datasets.

Conclusion

1 Multiple models reached AUC > 0.93, confirming feasibility.

2 KNN had the best accuracy; Random Forest had top precision.

3 Stratified K-Fold ensured robust validation.

4 Future work: use more diverse datasets, add clinical variables.

Thank You for Your Attention.