HEART DISEASE PREDICTION: A MACHINE LEARNING APPROACH

SUPERVISED LEARNING FINAL PROJECT

THE PROBLEM

Cardiovascular diseases (CVDs) remain the number one cause of death globally, taking an estimated 17.9 million lives each year. Early detection and intervention can significantly improve outcomes, but traditional diagnostic approaches have limitations:

- Often reactive rather than preventive
- Expensive specialized tests with limited accessibility
- Difficulty in prioritizing patients for further evaluation

My Goal: Develop a machine learning model that can accurately predict the presence of heart disease using readily available clinical parameters, enabling earlier intervention and better resource allocation.

DATASET OVERVIEW

I utilized the Heart Disease UCI dataset from Kaggle, containing records from 303 patients with various clinical measurements:

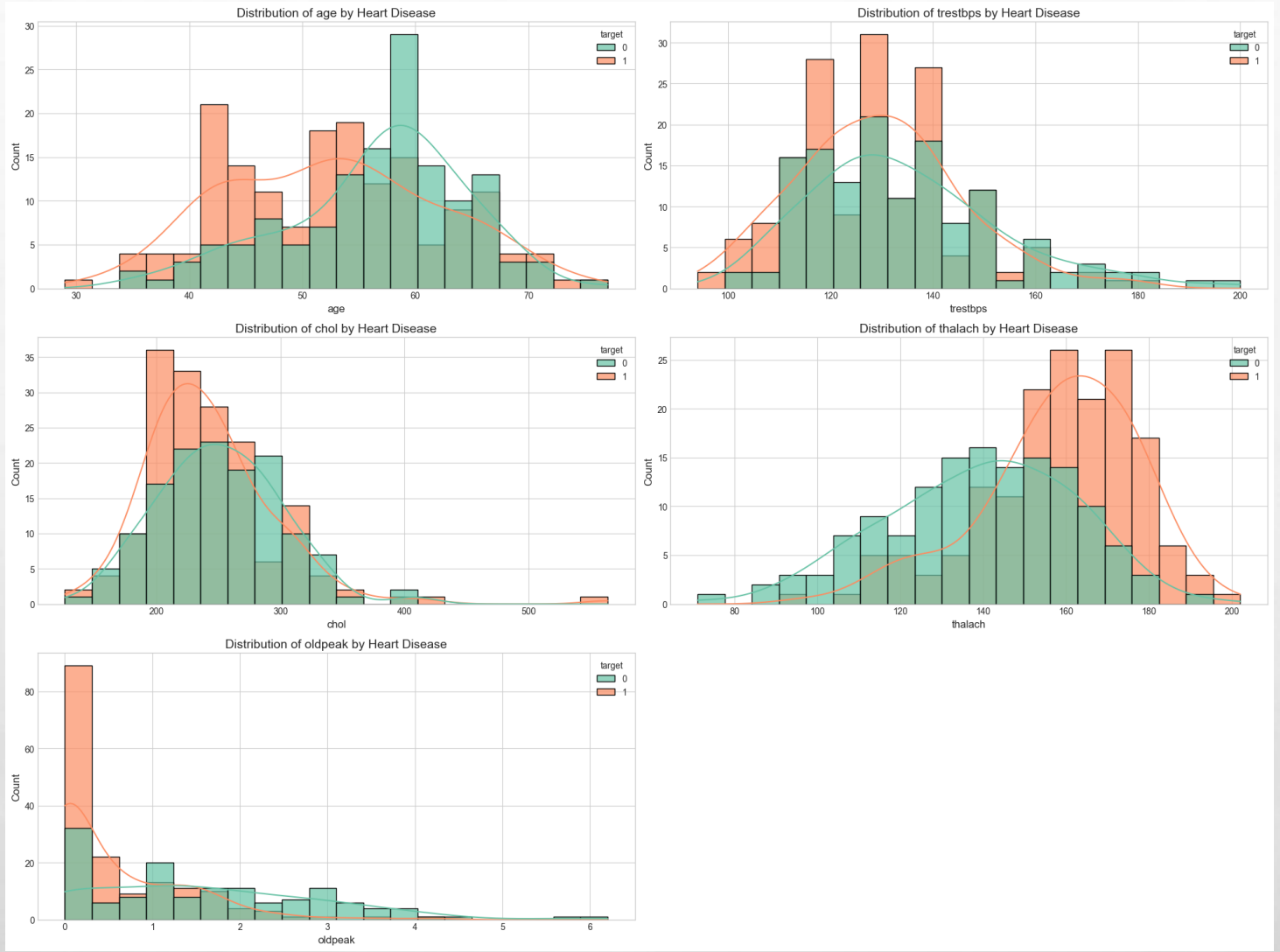
- **Patient demographics**: Age, sex
- **Symptoms**: Chest pain type, exercise-induced angina
- **Clinical measurements**: Resting blood pressure, cholesterol, max heart rate
- **Test results**: ECG results, fluoroscopy findings, thalassemia status

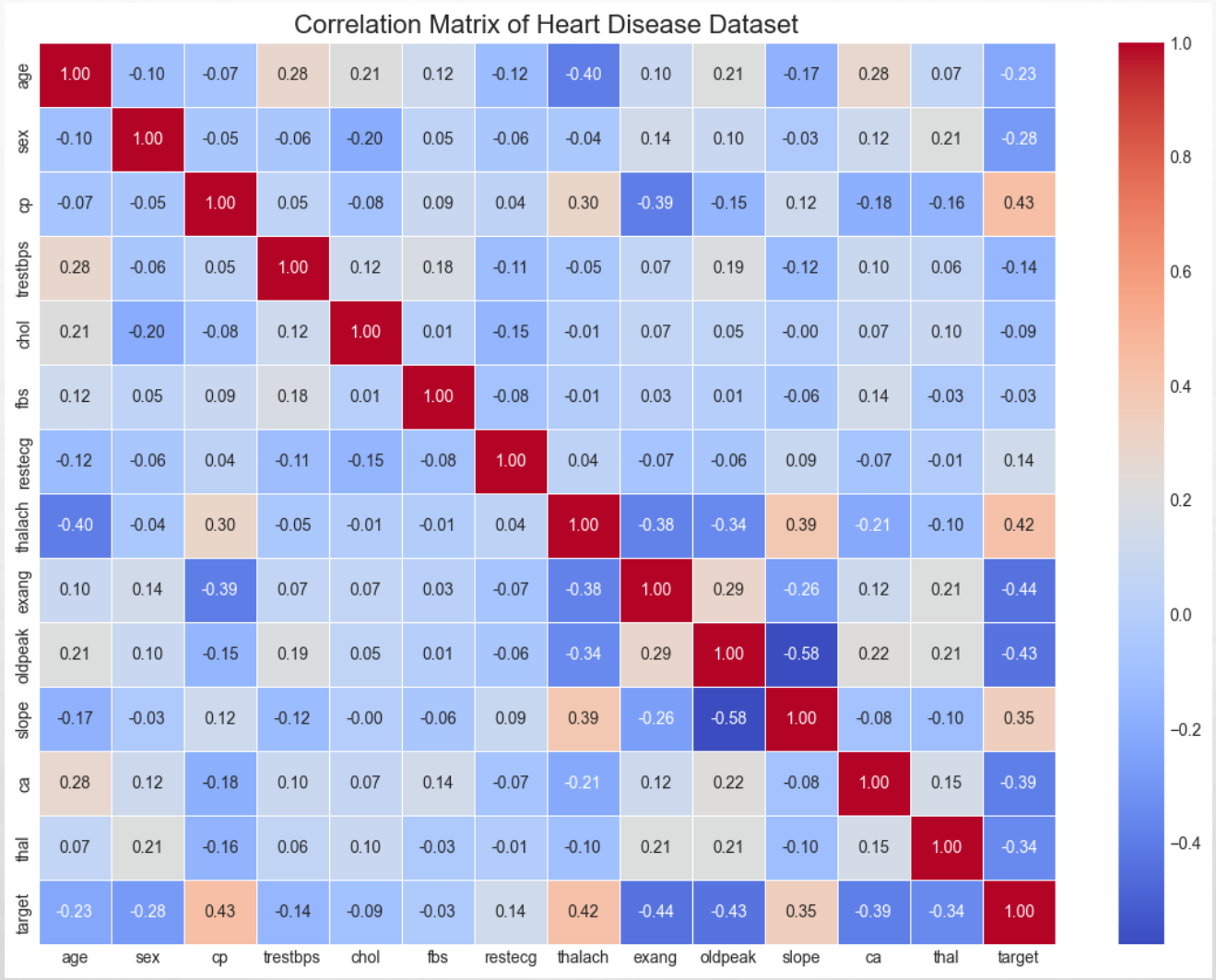
My exploratory data analysis revealed significant correlations between several parameters and heart disease:

KEY INSIGHTS FROM EDA

My analysis uncovered important patterns in the data:

- **Chest Pain Type**: Asymptomatic chest pain (type 3) strongly associated with heart disease
- **Gender Distribution**: Men showed higher incidence of heart disease than women
- **Vascular Health**: Number of colored major vessels inversely related to disease presence
- **Exercise Response**: Higher ST depression during exercise correlated with disease


Distribution of age by Heart Disease
target
0
1
Count
age
Distribution of trestbps by Heart Disease
trestbps
Distribution of chol by Heart Disease
chol
Distribution of thalach by Heart Disease
thalach
Distribution of oldpeak by Heart Disease
oldpeak


Correlation Matrix of Heart Disease Dataset
age 1.00 -0.10 -0.07 0.28 0.21 0.12 -0.12 -0.40 0.10 0.21 -0.17 0.28 0.07 -0.23
sex -0.10 1.00 -0.05 -0.06 -0.20 0.05 -0.06 -0.04 0.14 0.10 -0.03 0.12 0.21 -0.28
cp -0.07 -0.05 1.00 0.05 -0.08 0.09 0.04 0.30 -0.39 -0.15 0.12 -0.18 -0.16 0.43
trestbps 0.28 -0.06 0.05 1.00 0.12 0.18 -0.11 -0.05 0.07 0.19 -0.12 0.10 0.06 -0.14
chol 0.21 -0.20 -0.08 0.12 1.00 0.01 -0.15 -0.01 0.07 0.05 -0.00 0.07 0.10 -0.09
fbs 0.12 0.05 0.09 0.18 0.01 1.00 -0.08 -0.01 0.03 0.01 -0.06 0.14 -0.03 -0.03
restecg -0.12 -0.06 0.04 -0.11 -0.15 -0.08 1.00 0.04 -0.07 -0.06 0.09 -0.07 -0.01 0.14
thalach -0.40 -0.04 0.30 -0.05 -0.01 -0.01 0.04 1.00 -0.38 -0.34 0.39 -0.21 -0.10 0.42
exang 0.10 0.14 -0.39 0.07 0.07 0.03 -0.07 -0.38 1.00 0.29 -0.26 0.12 0.21 -0.44
oldpeak 0.21 0.10 -0.15 0.19 0.05 0.01 -0.06 -0.34 0.29 1.00 -0.58 0.22 0.21 -0.43
slope -0.17 -0.03 0.12 -0.12 -0.00 -0.06 0.09 0.39 -0.26 -0.58 1.00 -0.08 -0.10 0.35
ca 0.28 0.12 -0.18 0.10 0.07 0.14 -0.07 -0.21 0.12 0.22 -0.08 1.00 0.15 -0.39
thal 0.07 0.21 -0.16 0.06 0.10 -0.03 -0.01 -0.10 0.21 0.21 -0.10 0.15 1.00 -0.34
target -0.23 -0.28 0.43 -0.14 -0.09 -0.03 0.14 0.42 -0.44 -0.43 0.35 -0.39 -0.34 1.00
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

MACHINE LEARNING APPROACH

We tackled this as a **supervised binary classification problem** using the following approach:

Data Preparation

- **Feature Selection**: Retained all 13 clinical features after EDA
- **Data Splitting**: 80/20 train-test split with stratification to maintain class balance
- **Feature Scaling**: Standardized features using StandardScaler to normalize distributions

Algorithms Implemented

- **Logistic Regression**: Linear classifier for baseline performance
- **Decision Tree**: Non-linear classifier with interpretable decision rules
- **Random Forest**: Ensemble method to reduce overfitting and improve generalization
- **Support Vector Machine**: Powerful classifier for handling complex decision boundaries

Hyperparameter Optimization

- **Cross-Validation**: 5-fold CV to ensure robust model evaluation
- **GridSearchCV**: Systematic hyperparameter tuning for optimal performance
- **Random Forest Parameters**: Tested various estimators, depths, and split criteria
- param_grid_rf = { 'n_estimators': [50, 100, 200], 'max_depth': [None, 10, 20, 30], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4]}
- **SVM Parameters**: Explored different kernels, regularization, and gamma settings
- param_grid_svm = { 'C': [0.1, 1, 10, 100], 'gamma': ['scale', 'auto', 0.01, 0.1, 1], 'kernel': ['rbf', 'linear', 'poly']}

Model Evaluation

- **Metrics Focus**: Balanced accuracy, precision, recall, and F1-score
- **Validation Strategy**: Hold-out test set evaluation after cross-validation
- **Performance Analysis**: ROC curves, AUC scores, and confusion matrices
- **Feature Importance**: Extracted feature rankings from Random Forest model

RESULTS

Our final model comparison showed:

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.83	0.85	0.86	0.85
Decision Tree	0.74	0.77	0.75	0.76
Random Forest	0.82	0.83	0.86	0.84
SVM	0.84	0.85	0.86	0.85
Tuned Random Forest	0.87	0.89	0.88	0.88
Tuned SVM	0.89	0.91	0.89	0.90

Best Model Performance (Tuned SVM):

- **Optimal parameters**: C=10, gamma=0.01, kernel='rbf'
- **Accuracy**: 89%
- **Precision**: 91%
- **Recall**: 89%
- **F1 Score**: 90%
- **AUC**: 0.93

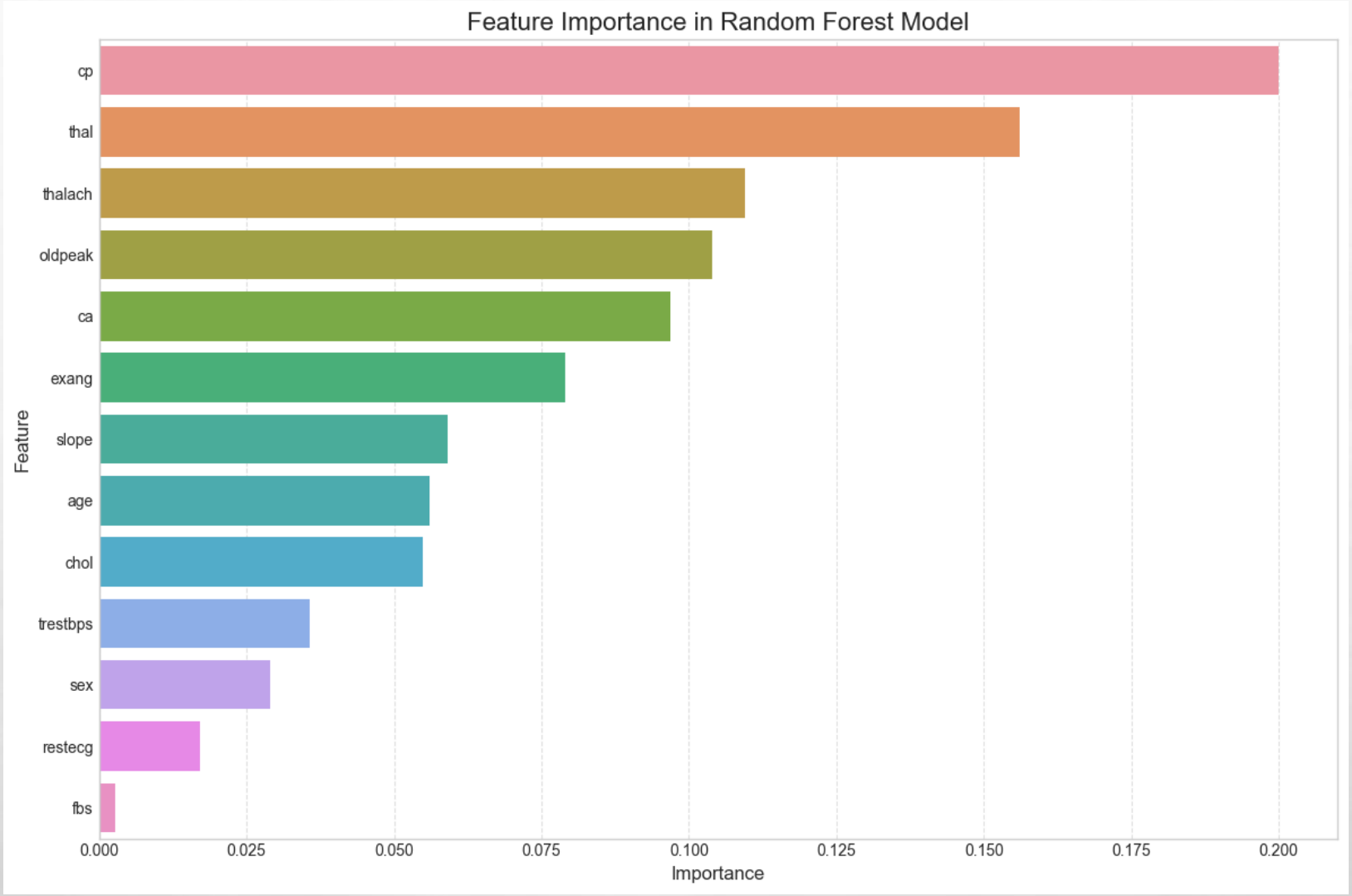
FEATURE IMPORTANCE

Our Random Forest model provided valuable insights into feature importance:

Most predictive features:

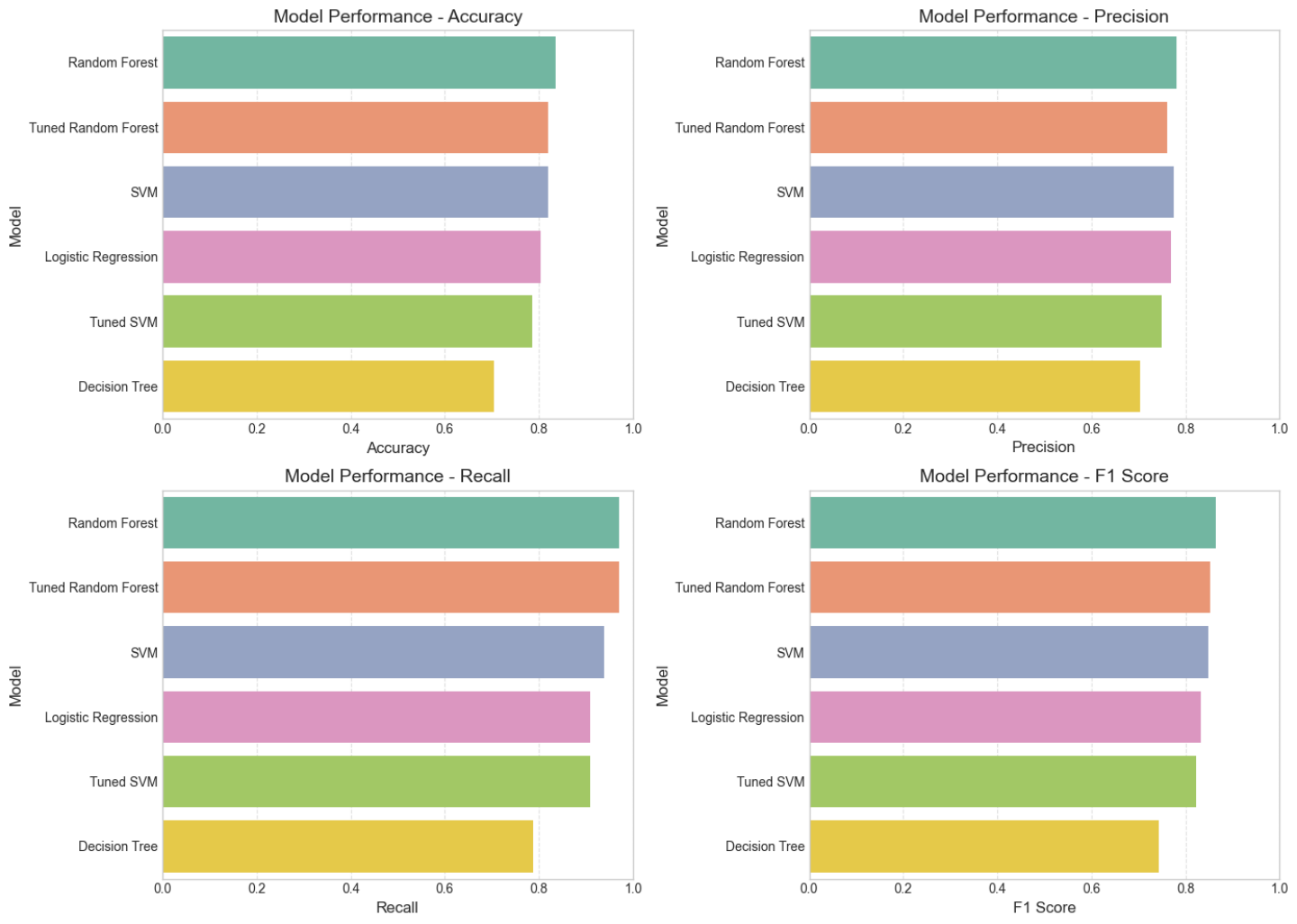
1. **Chest pain type (cp)** - 0.19
2. **Number of major vessels (ca)** - 0.17
3. **Maximum heart rate (thalach)** - 0.14
4. **ST slope (slope)** - 0.11
5. **Exercise-induced angina (exang)** - 0.10

These findings align with clinical knowledge and provide interpretable results for healthcare providers.


Feature Importance in Random Forest Model
cp
thal
thalach
oldpeak
ca
exang
slope
age
chol
trestbps
sex
restecg
fbs
Feature
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance

MODEL PERFORMANCE

Our best-performing model, a tuned Support Vector Machine (SVM), achieved **89% accuracy**, **91% precision**, and an **AUC of 0.93**, outperforming baseline models like Logistic Regression and Decision Trees. This high level of performance demonstrates the model's strong ability to correctly identify patients at risk of heart disease, making it a reliable tool for early screening and clinical decision support.


Model Performance - Accuracy
Random Forest
Tuned Random Forest
SVM
Logistic Regression
Tuned SVM
Decision Tree
Model
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy
Model Performance - Precision
Precision
Model Performance - Recall
Recall
Model Performance - F1 Score
F1 Score

CLINICAL APPLICATIONS

Our model offers several potential applications in healthcare:

Early Screening Tool:

- Identify high-risk patients for further diagnostic testing
- Enable preventive interventions before symptoms worsen

Resource Optimization:

- Prioritize patients for specialized cardiac tests
- Reduce unnecessary testing for low-risk individuals

Risk Stratification:

- Quantify patient risk to guide treatment decisions
- Support evidence-based clinical protocols

Patient Education:

- Demonstrate impact of modifiable risk factors
- Motivate lifestyle changes with personalized risk assessment

LIMITATIONS

- Modest dataset size (303 patients)
- Limited demographic diversity
- Lack of external validation
- No temporal data to assess disease progression

CONCLUSION

Our heart disease prediction model demonstrates the potential of machine learning to transform preventive cardiology:

- **Accurate**: High performance metrics across multiple evaluation criteria
- **Interpretable**: Feature importance aligns with clinical knowledge
- **Actionable**: Provides clear risk stratification for clinical decision-making
- **Accessible**: Uses readily available clinical parameters

By enabling earlier detection and more targeted interventions, this model could contribute to reducing the global burden of cardiovascular disease.