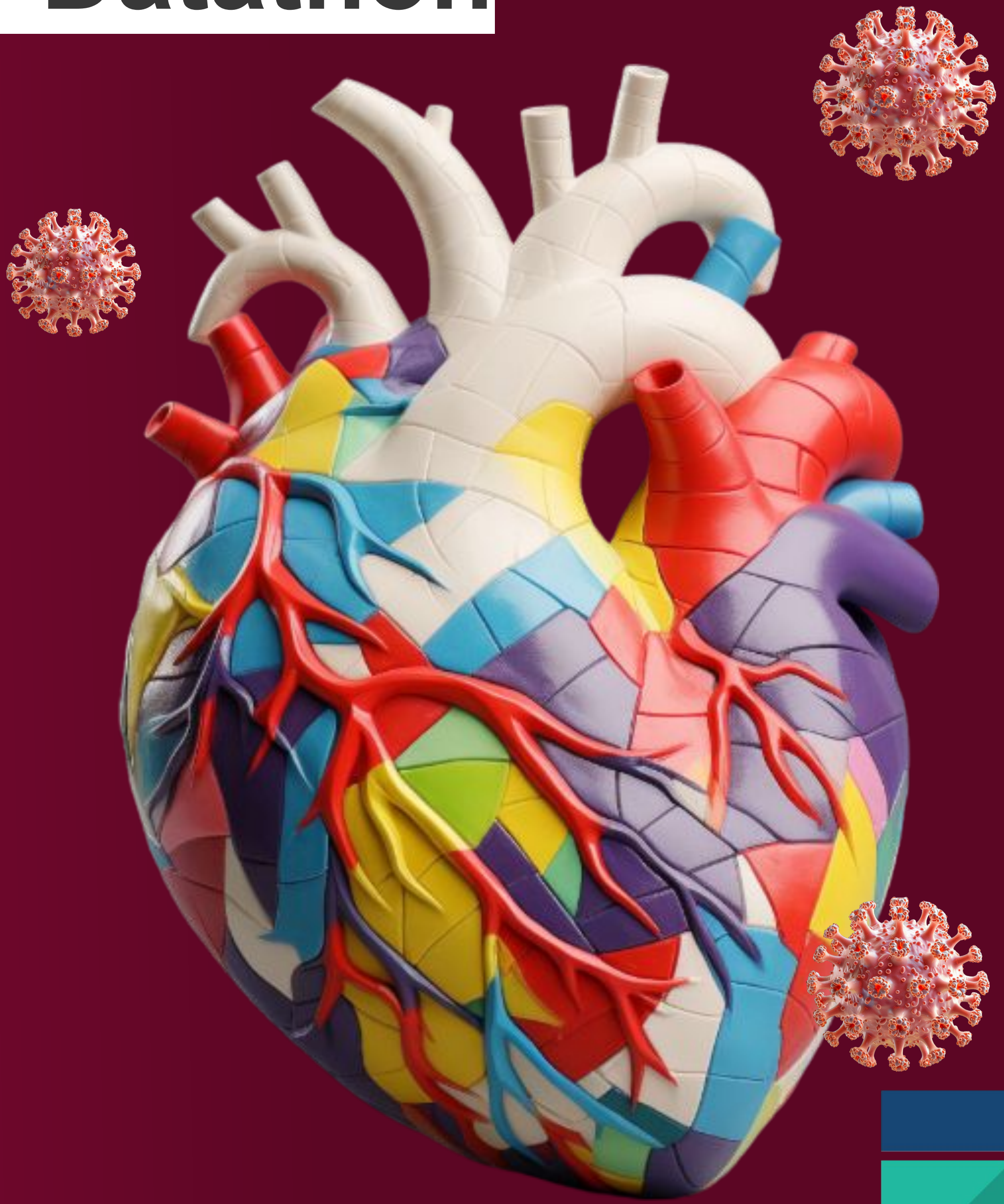
Kshitij 2025 x Altair Datathon

Heart Disease Prediction



Theme: Health
Problem Statement: Heart Disease Diagnosis: Predictive Modeling with Patient Health Data

Prepared By: Abhinav Saxena
Platform Used: Altair AI Studio

Project Overview

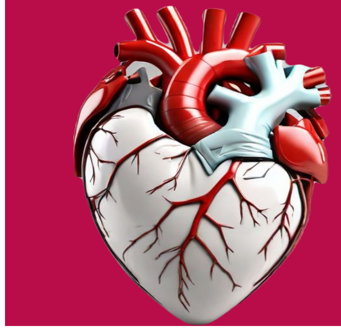
This project aims to predict the likelihood of heart disease using key clinical indicators such as age, cholesterol levels, blood pressure, and other factors.

Approach

The predictive model was built using machine learning algorithms in **Altair AI Studio**, leveraging data preprocessing and feature engineering.

Dataset Overview : The dataset contains clinical health data from Kaggle, used to predict heart disease.

kaggle

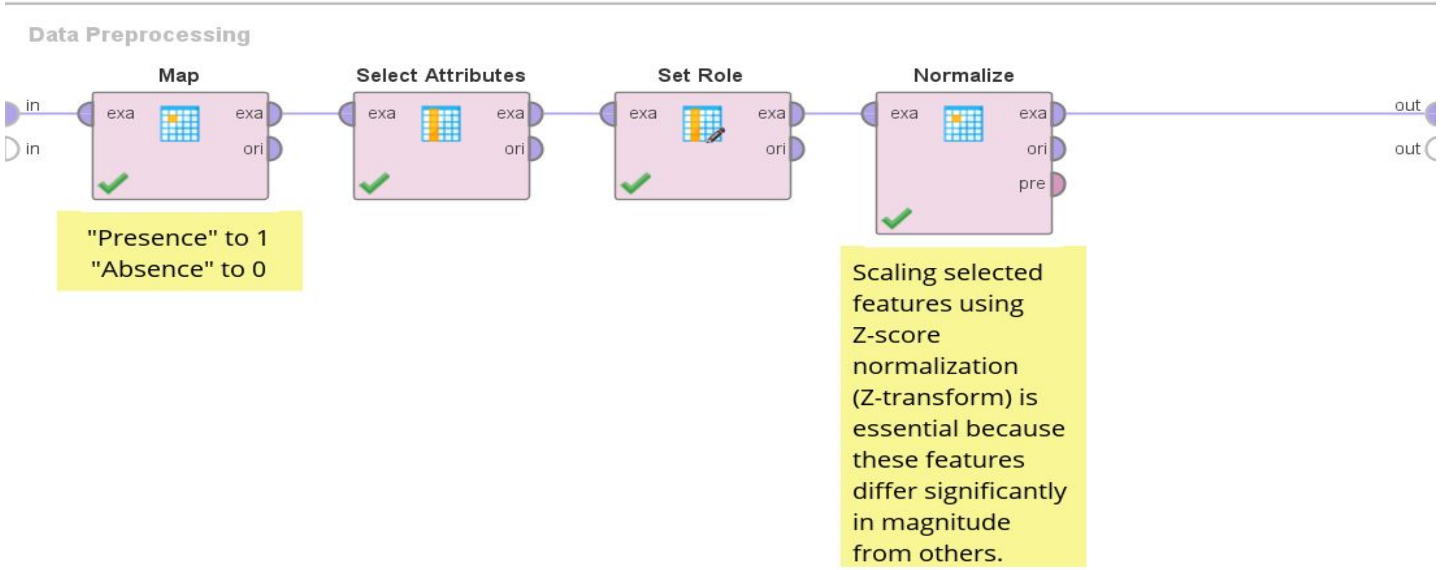


Age: Patient's age
Sex: Gender (0 = Female, 1 = Male)
Chest Pain Type: Type of chest pain (4 types)
BP: Blood Pressure
Cholesterol: Serum cholesterol level in mg/dl
FBS over 120: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
EKG Results: Electrocardiographic results (values range from 0 to 2)
Max HR: Maximum heart rate achieved
Exercise Angina: Whether the patient experiences angina during exercise (1 = yes, 0 = no)
ST Depression: Depression of the ST segment during exercise (continuous value)
Slope of ST: The slope of the ST segment (3 types)
Number of Vessels Fluoroscopy: Number of vessels colored by fluoroscopy (0 to 3)
Thallium: Thallium stress test result (values 3, 6, 7)
Heart Disease: Presence or absence of heart disease (Presence, Absence)

Data Preprocessing

Data preprocessing is critical to ensure clean, consistent, and usable data for accurate model predictions.

Missing Values: No Missing Values found.
Encoding Categorical Variables : Encoding Target Variable (Heart Disease) using Map Operator.
Set Role: Setting Target Variable using Set Role Operator.
Scaling: Z-Transform using Normalize Operator.



"Presence" to 1
"Absence" to 0

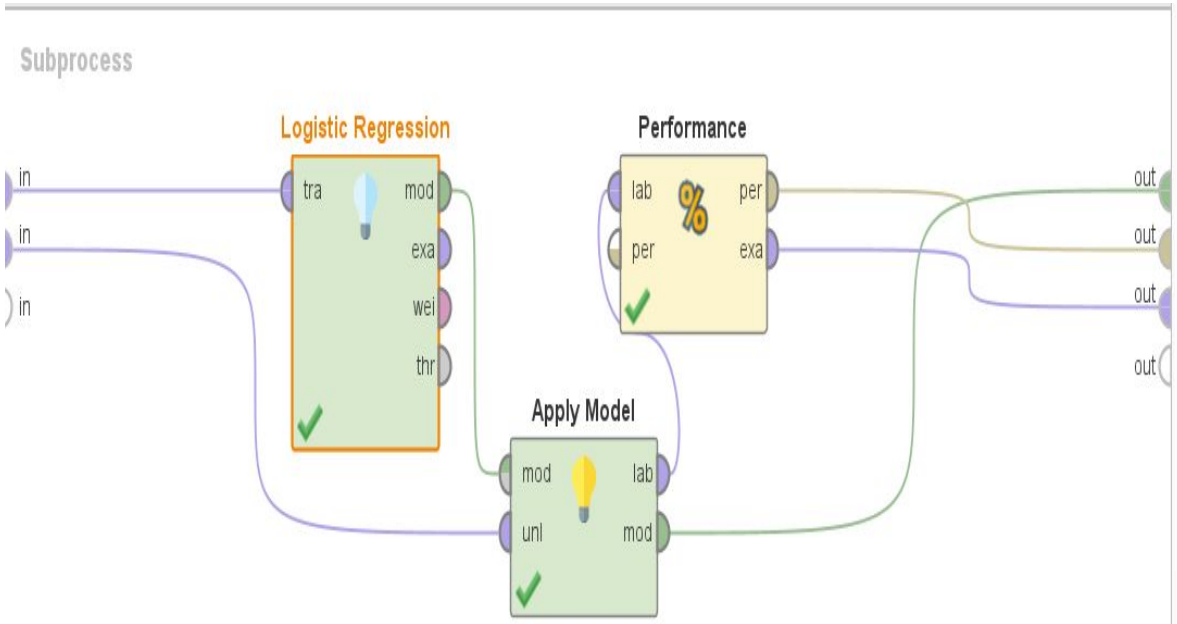
Scaling selected features using Z-score normalization (Z-transform) is essential because these features differ significantly in magnitude from others.

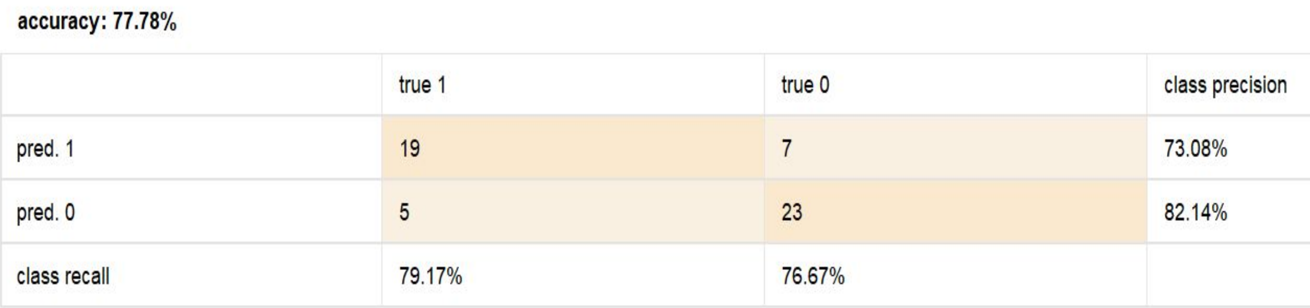
Baseline Model: Logistic Regression

Data Splitting: The dataset was split into 80% for training and 20% for testing to evaluate model performance.

Logistic Regression was chosen as the baseline model due to its simplicity and effectiveness in binary classification problems like heart disease prediction.

The model achieved an accuracy of 77.78%."





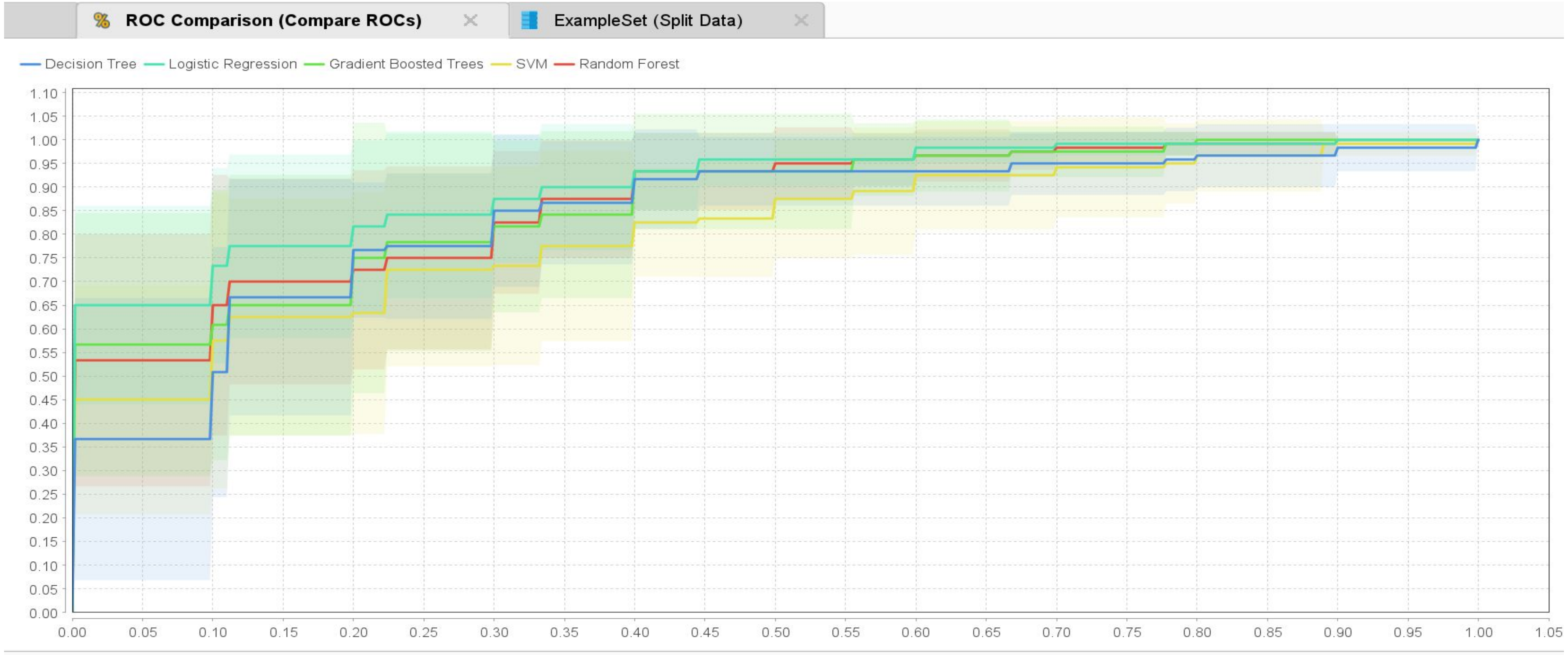
accuracy: 77.78%

	true 1	true 0	class precision
pred. 1	19	7	73.08%
pred. 0	5	23	82.14%
class recall	79.17%	76.67%	

Comparing ROC

ROC of 5 models - Logistic Regression, Decision Tree, Random Forest, SVM, Gradient Boosted Trees were compared using Compare ROC Operator

Logistic Regression achieved the highest ROC AUC score, indicating strong predictive power for heart disease diagnosis.



Model Selection and Optimization

CV: 5 Folds

Logistic Regression

A linear model suitable for binary classification, chosen as the baseline.

Decision Trees

Splits data into branches for classification or regression.

SVM

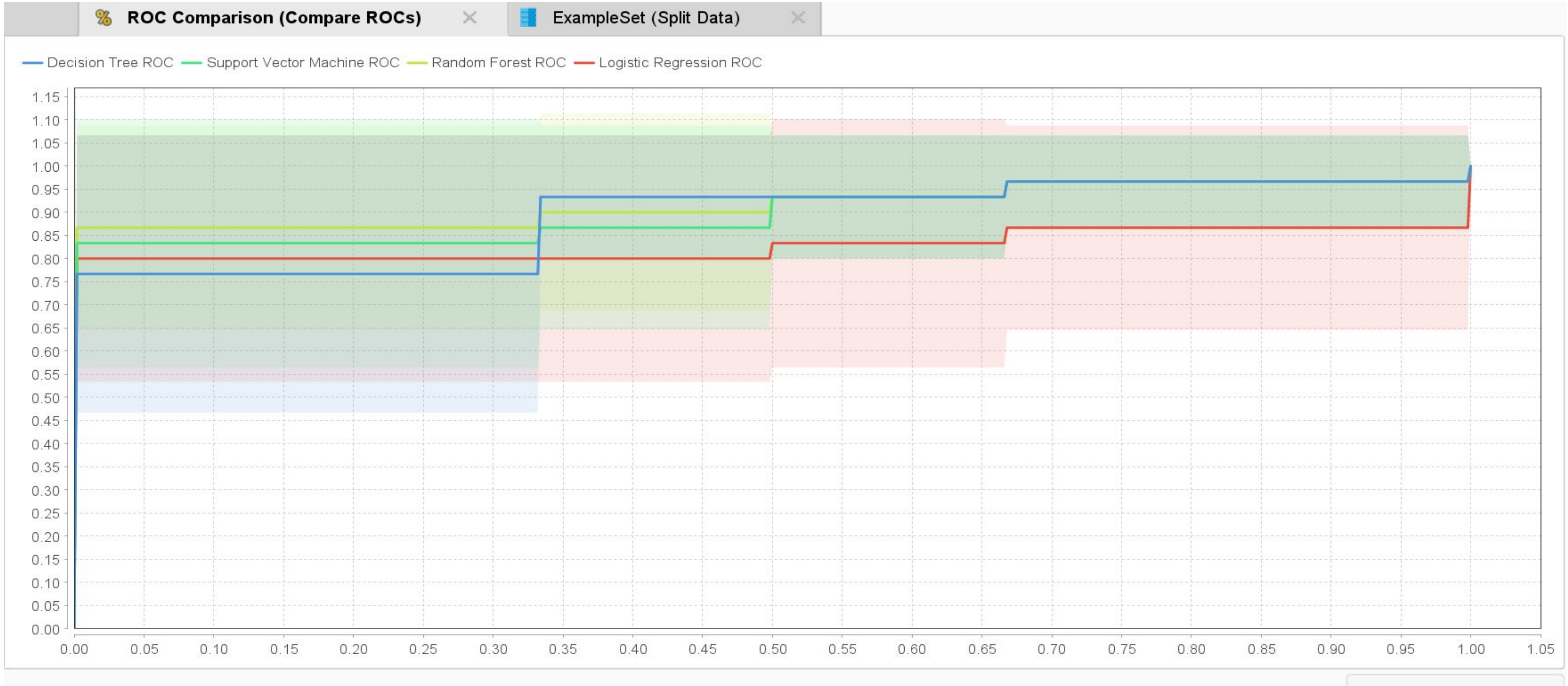
Finds the optimal hyperplane to classify data points effectively.

Random Forest

An ensemble of decision trees that improves accuracy and reduces overfitting..

Comparing ROC with Optimized Parameterized and CV of 5 folds

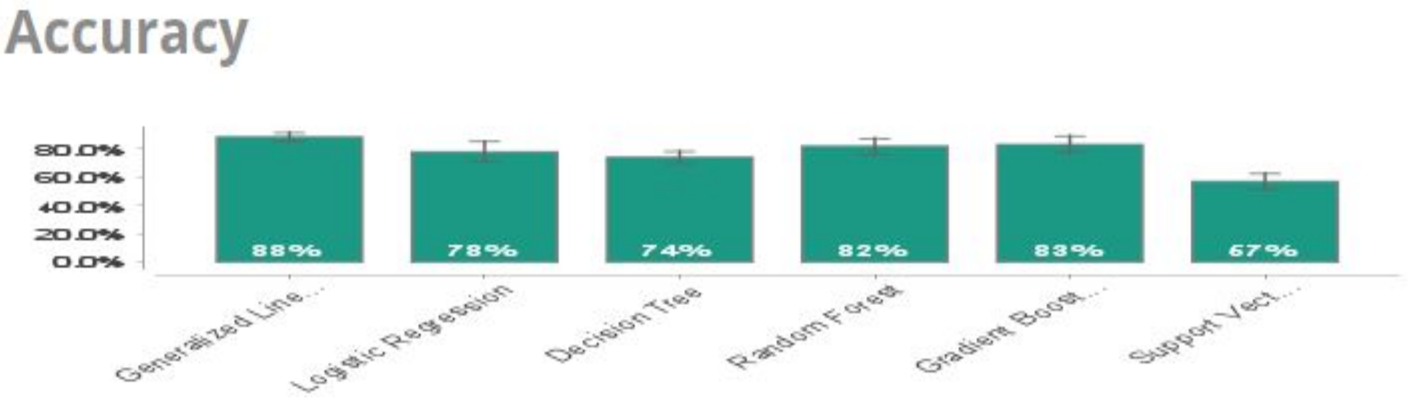
Decision Tree achieved the highest ROC AUC score, indicating strong predictive power for heart disease diagnosis.

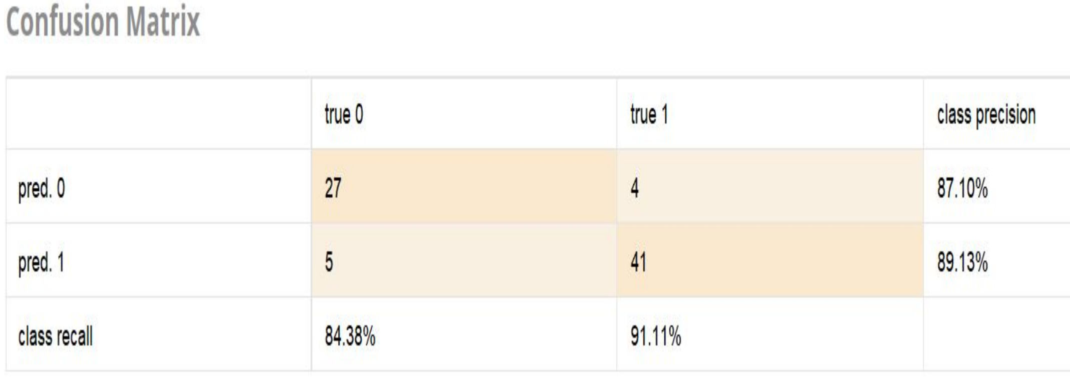


Auto Model Selection

1. Generalized Linear Model
2. Logistic Regression
3. Decision Tree
4. Random Forest
5. Gradient Boosted Trees
6. Support Vector Machine

Generalized Linear Model performed the best with Accuracy of 88.3%
Gradient Boosted Trees also performed well with Accuracy of 83.1%





Confusion Matrix

	true 0	true 1	class precision
pred. 0	27	4	87.10%
pred. 1	5	41	89.13%
class recall	84.38%	91.11%	

Results

Key clinical indicators in predicting heart disease include Max Heart Rate, ST Depression, and Vessels Fluoroscopy. High Max Heart Rate during physical activity, significant ST Depression in ECG readings, and an increased number of affected vessels on fluoroscopy are strongly associated with higher risk.

In our model, these factors emerged as the most predictive features, highlighting their critical role in accurately identifying patients at risk for heart disease."

Learnings

Through Altair AI Studio and RapidMiner, I gained hands-on experience in data preprocessing, model selection, and performance evaluation. Altair AI Studio enabled efficient data visualization and model building, while RapidMiner provided an intuitive workflow for applying machine learning algorithms and fine-tuning models.
This process enhanced my understanding of end-to-end machine learning pipelines and the importance of iterative optimization in achieving accurate predictions."

Thank You