Heart Attack Risk Prediction: Capstone 2 Project

Presented by: Nilesh Suresh

Context

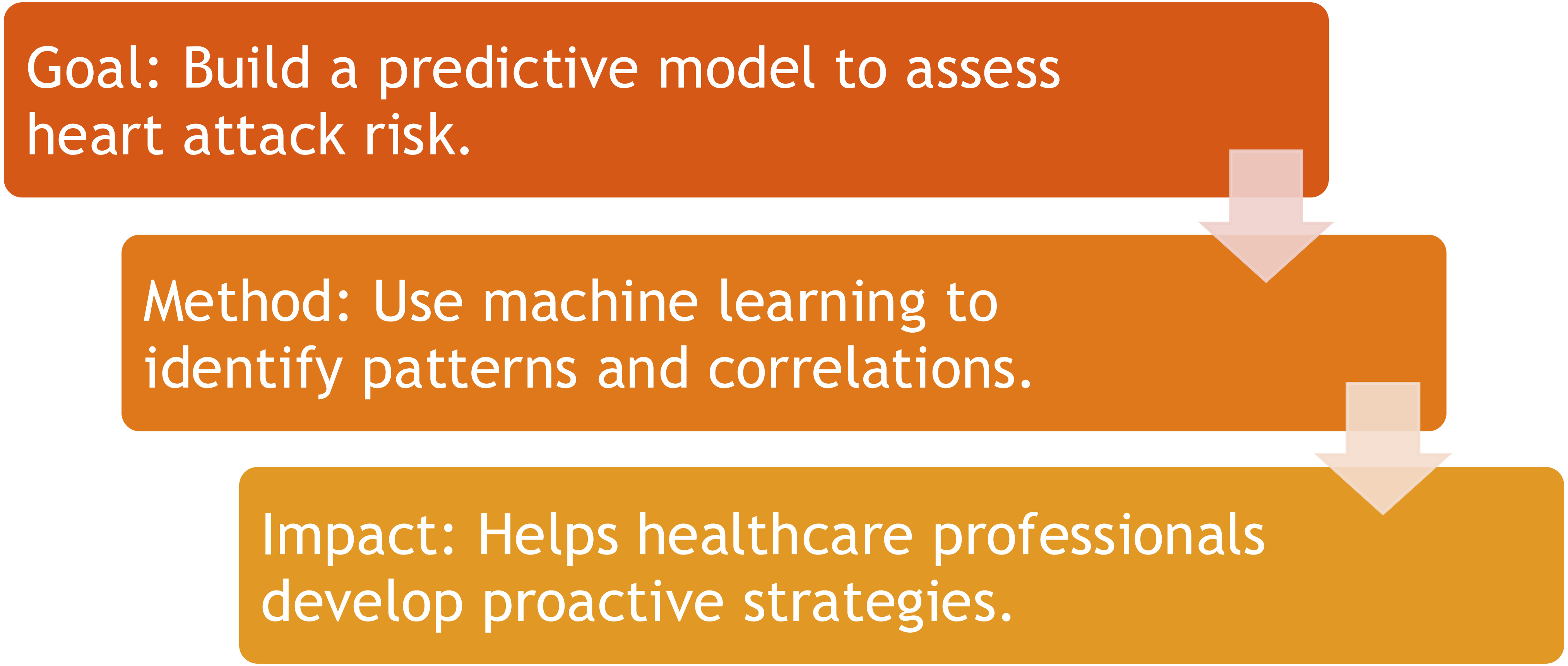
Heart attacks are a major global health issue.

Understanding risk factors can help prevent them.

The dataset includes key health attributes (age, cholesterol, blood pressure, lifestyle habits, etc.).

Using machine learning, we aim to create predictive models for better prevention and management.

Problem Statement



Goal: Build a predictive model to assess heart attack risk.

Method: Use machine learning to identify patterns and correlations.

Impact: Helps healthcare professionals develop proactive strategies.

Data Wrangling

Initial dataset: 8,763 rows, 25 columns.

Cleaning steps:

- Rounding values for consistency.
- Splitting blood pressure into Systolic & Diastolic.

Outcome: A well-structured dataset for further analysis.

Data Features

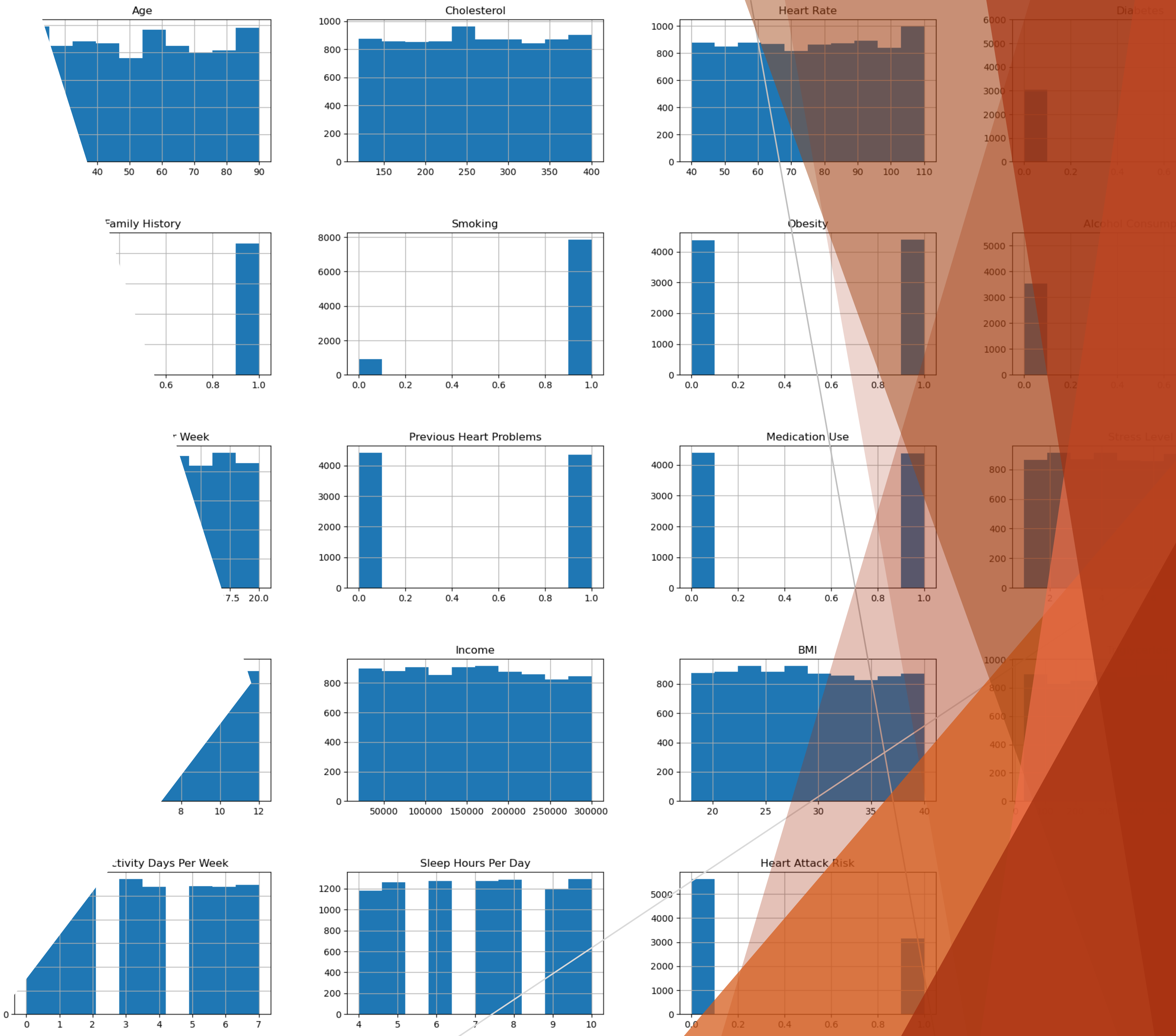
- Numeric Features
  - Age
  - Cholesterol
  - Heart Rate
  - Exercise Hours Per Week
  - Stress Level
  - Sedentary Hours Per Day
  - Income
  - BMI
  - Triglycerides
  - Physical Activity Days Per Week
  - Sleep Hours Per Day
  - Diabetes
  - Family History
  - Smoking
  - Obesity
  - Alcohol Consumption
  - Previous Heart Problems

- Categorical Features
  - Blood Pressure
  - Sex
  - Diet
  - Country
  - Continent
  - Hemisphere

- Target Feature
  - Heart Attack Risk

Distributions Of Feature Values

Distribution analyses were conducted for both numerical and categorical variables, confirming that the feature distributions were appropriate for heart attack risk prediction modeling.



Exploratory Data Analysis (EDA)

Purpose: Identify trends, anomalies, and relationships between variables.

Data transformations:

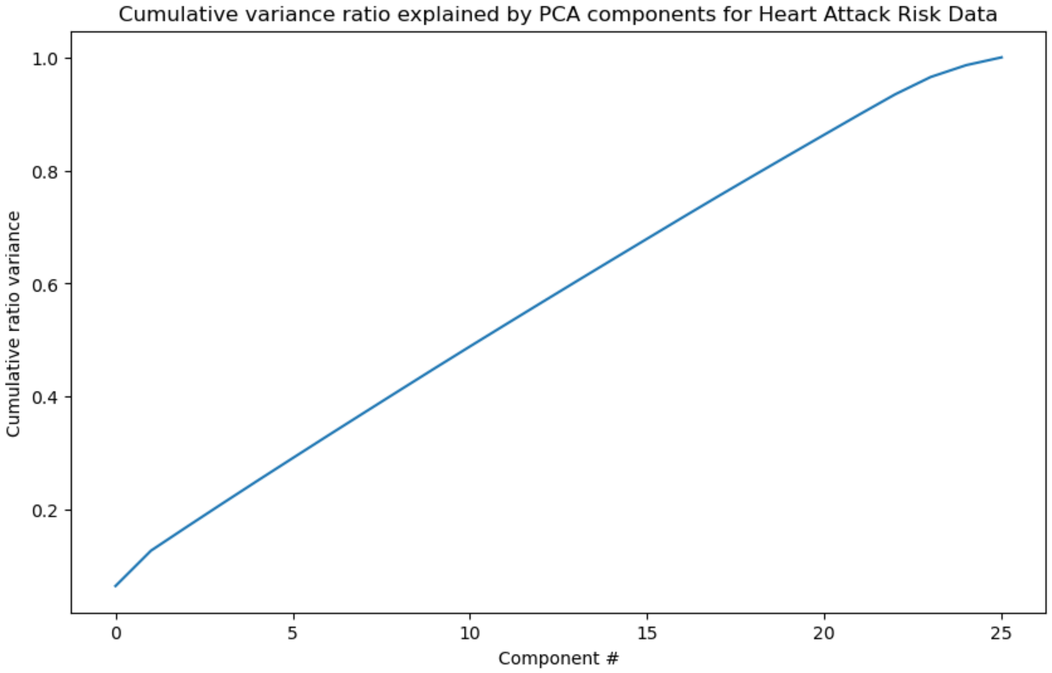
- Label encoding categorical variables.
- Feature mapping (Cholesterol & Blood Pressure categories).

Key positively correlated features:

Cholesterol, Diabetes, Exercise, Triglycerides, Blood Pressure, Age.

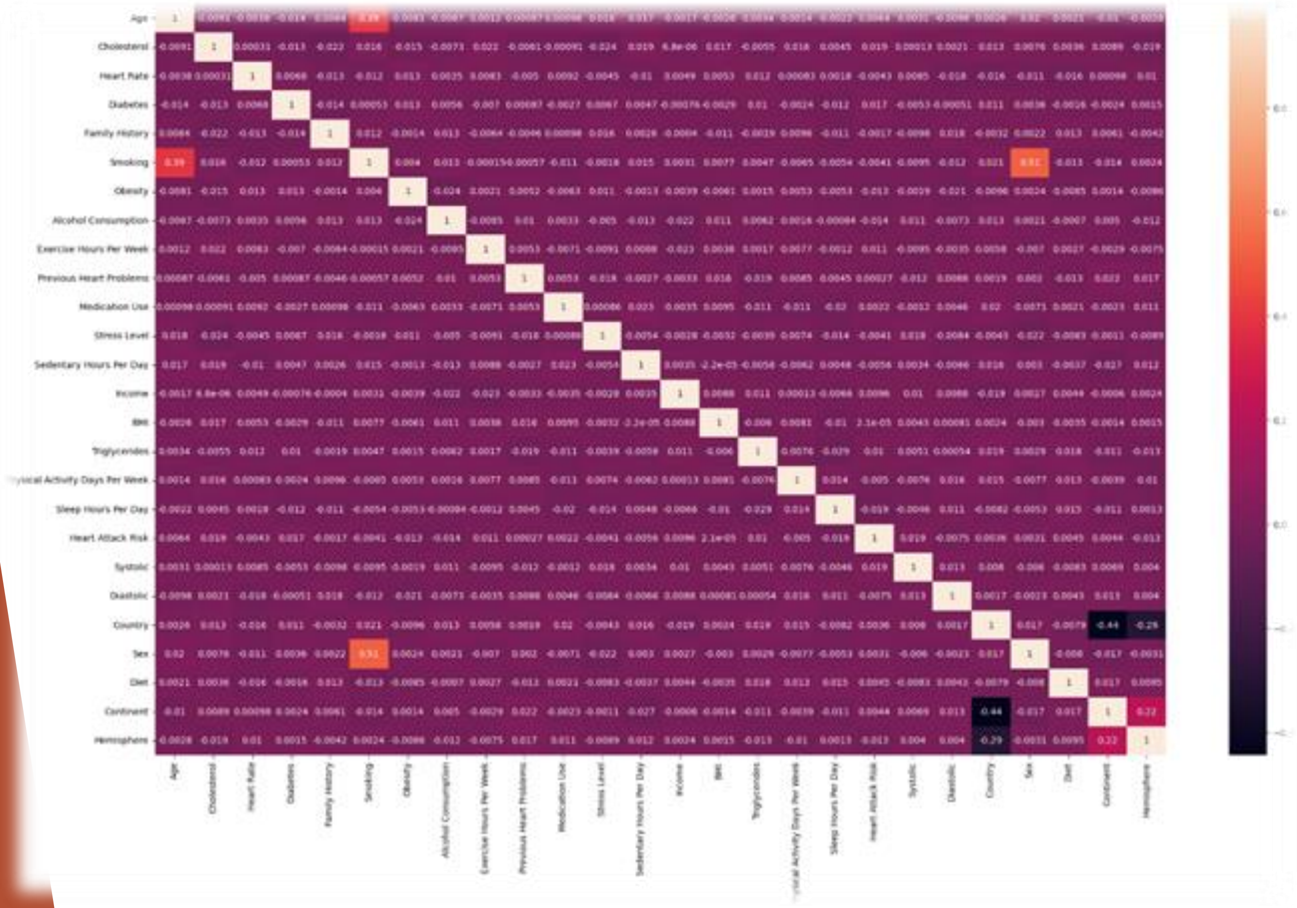
Findings: Older patients with high cholesterol/blood pressure have higher heart attack risk.

Principal Component Analysis (PCA)



- Standardized the dataset using scaling to ensure uniformity across features.
- PCA transformation was applied to capture variance in the dataset.
- The first principal component (Age) explained 15% variance, with about 80% variance coming from the top 15 components

Correlation Analysis



- Heatmap visualization revealed strong positive correlations with "Heart Attack Risk", particularly:
  - Cholesterol
  - Diabetes
  - Exercise Hours Per Week
  - Triglycerides
  - Systolic Blood Pressure
  - Age
  - Previous Heart Problems
  - Medication Use

Data Preprocessing

- Dummy feature creation: Converted categorical variables into numeric format.
- Encoding categorical features (Sex, Diet, Country, Continent, Hemisphere).
- Train/Test Split: 80% training, 20% testing.
- Outcome: Data ready for machine learning model development.

Modeling Approach

Machine Learning Models Tested:

Logistic Regression

Random Forest

K-Nearest Neighbors (KNN)

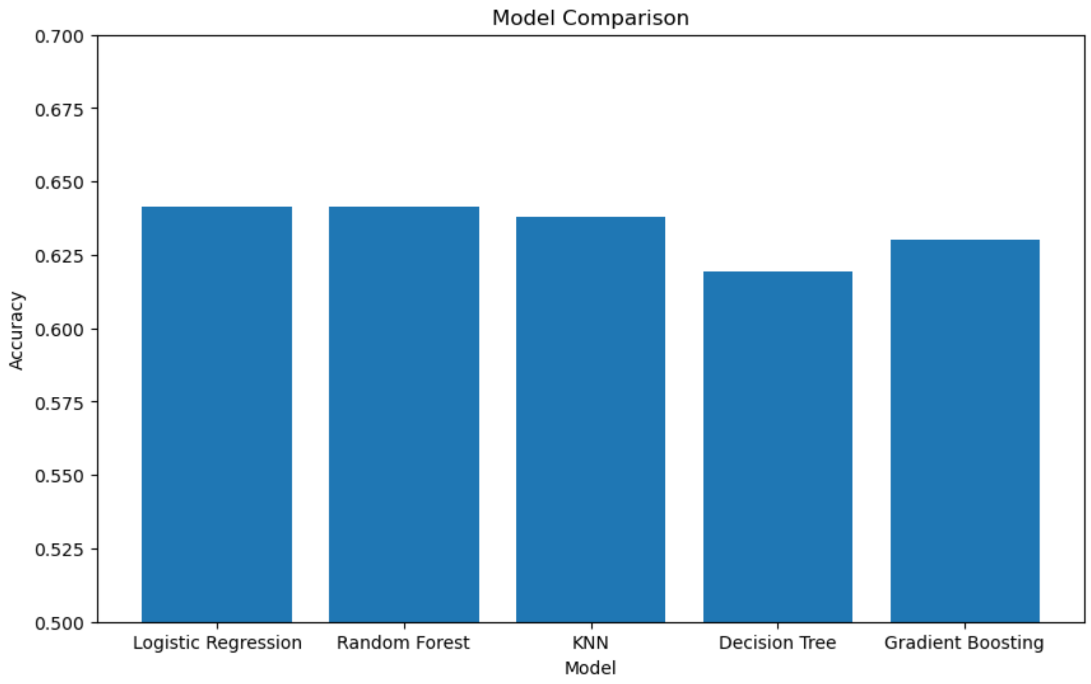
Decision Tree

Gradient Boosting

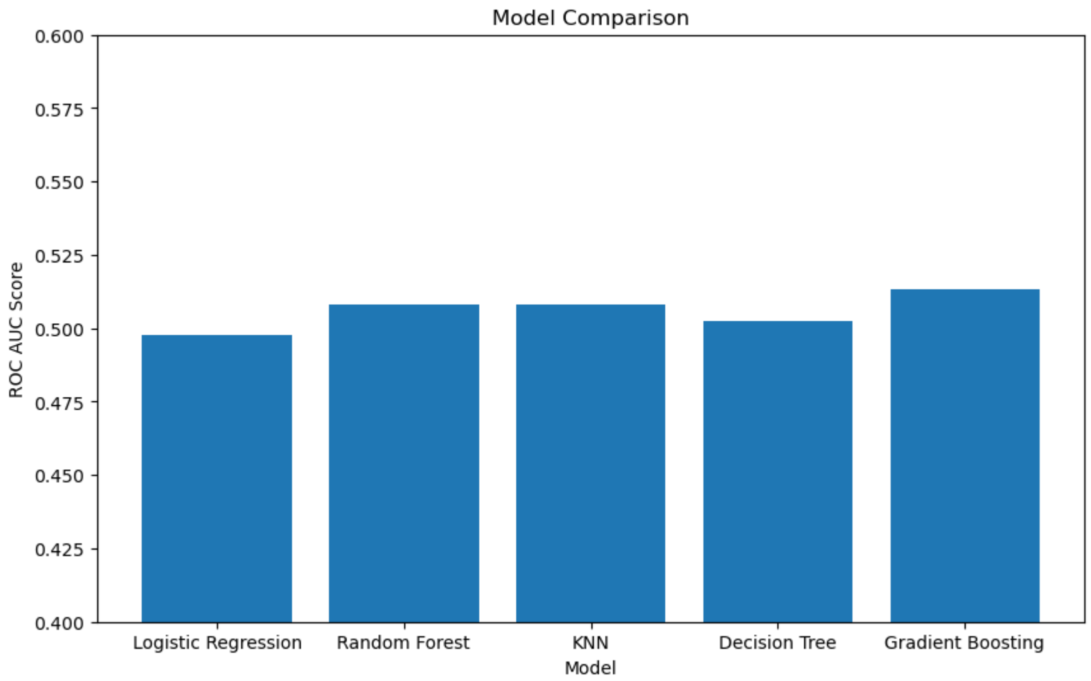
Key Evaluation Metrics: Accuracy, ROC AUC Score, Confusion Matrix.

Model Evaluation

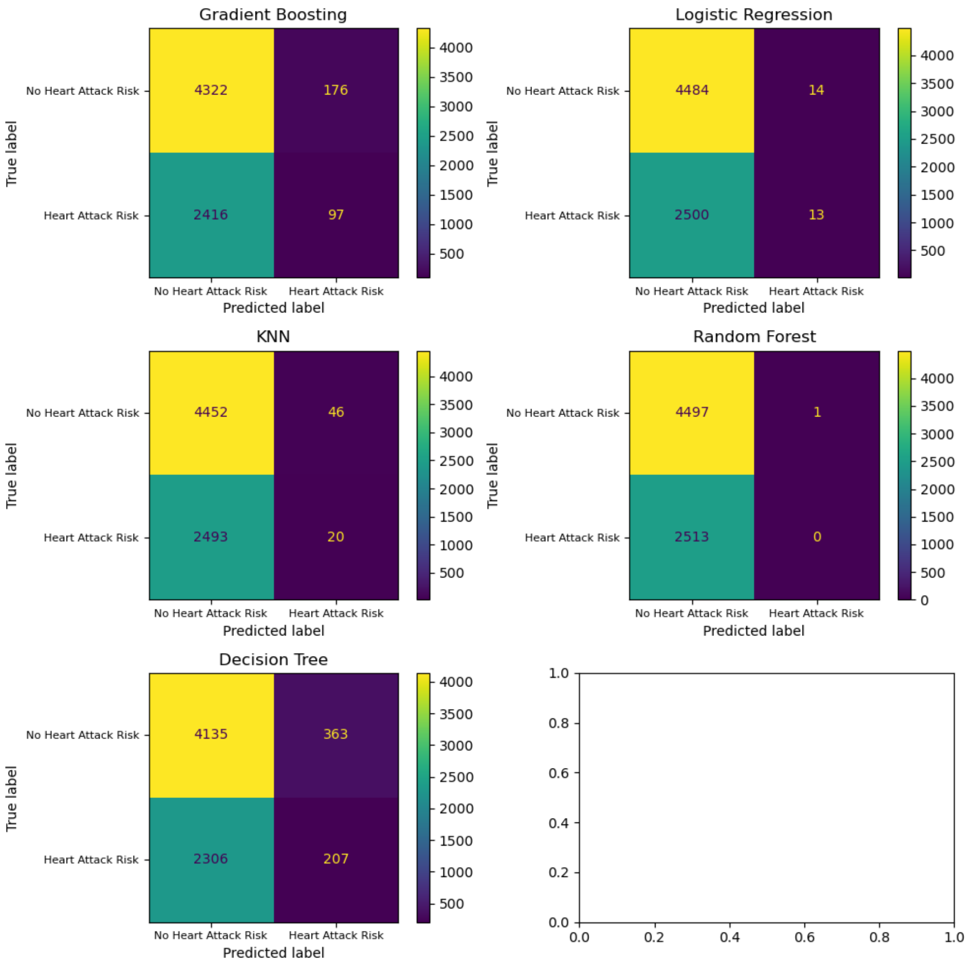
Classifier	Best Hyperparameters	Accuracy	Confusion Matrix	ROC AUC Score
Logistic Regression	C = 0.2113	64.10%	Correctly predicted: 4484 No-Risk, 13 Risk cases. Misclassified: 2500 No-Risk as Risk, 14 Risk as No-Risk.	0.4978 (poor class separation)
Random Forest Classifier	Criterion: gini, Max Depth: 3, Min Samples Split: 2, N Estimators: 200	64.10%	Correctly predicted: 4497 No-Risk, 0 Risk cases. Misclassified: 2513 No-Risk as Risk, 1 Risk as No-Risk.	0.5079 (slightly better but still weak)
KNeighbors Classifier	n_neighbors = 47	63.80%	Correctly predicted: 4452 No-Risk, 20 Risk cases. Misclassified: 2493 No-Risk as Risk, 46 Risk as No-Risk.	0.5070 (similar to Random Forest)
Decision Tree Classifier	Criterion: entropy, Max Depth: 3, Min Samples Split: 2	61.90%	Correctly predicted: 4135 No-Risk, 207 Risk cases. Misclassified: 2306 No-Risk as Risk, 363 Risk as No-Risk.	0.5027 (weak class separation)
Gradient Boosting Classifier	Max Depth: 3, Min Samples Split: 2, Learning Rate: 0.1	63.00%	Correctly predicted: 4322 No-Risk, 97 Risk cases. Misclassified: 2416 No-Risk as Risk, 176 Risk as No-Risk.	0.5134 (slightly better, but weak)



Model Comparison - Accuracy



Model Comparison – ROC AUC Score



Model Comparison – Confusion Matrix

Model Comparisons & Insights

Best-performing models: Gradient Boosting and Decision Tree.

Key observations:

- Strong predictors: Cholesterol, Triglycerides, Blood Pressure, Age, Previous Heart Problems.
- High cholesterol & sedentary lifestyle contribute significantly to heart attack risk.

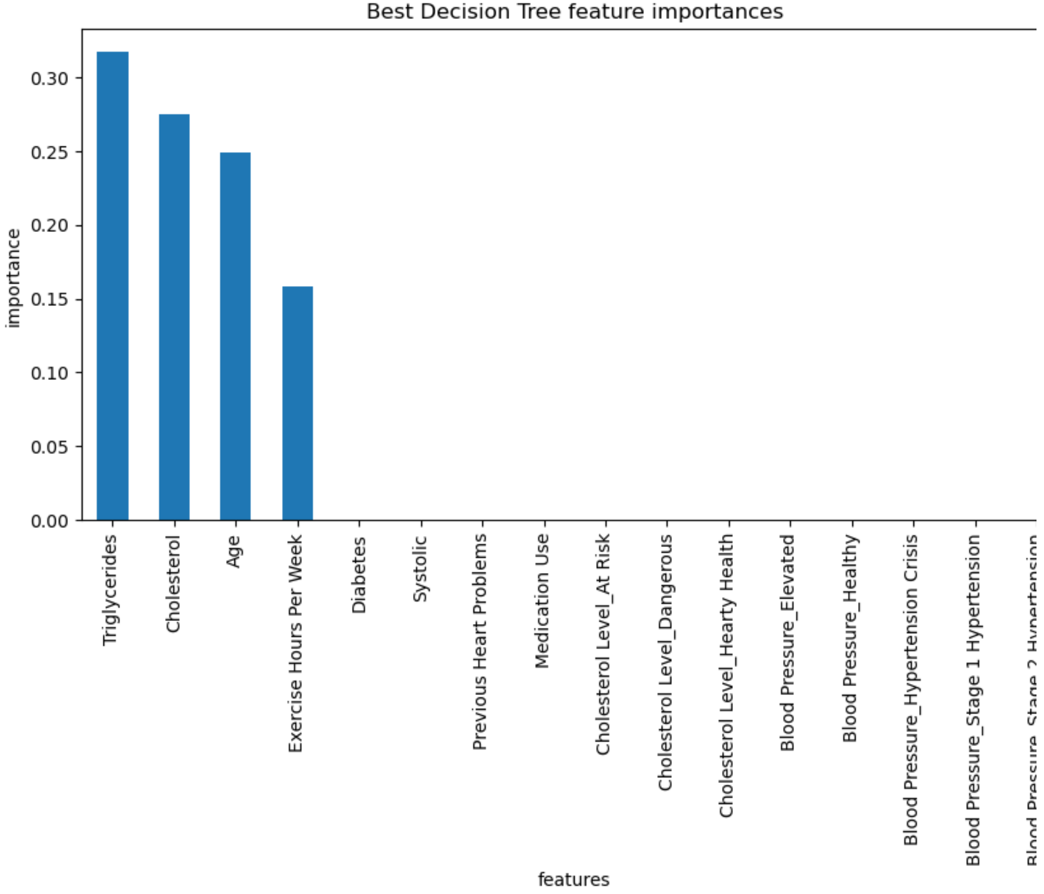
Final Model Selection

Chosen model: Decision Tree Classifier (best at identifying heart attack risk).

Feature Importance:

Top predictors include Triglycerides, Cholesterol, Age, and Exercise.

Scenario Analysis: Simulated risk under different lifestyle conditions.

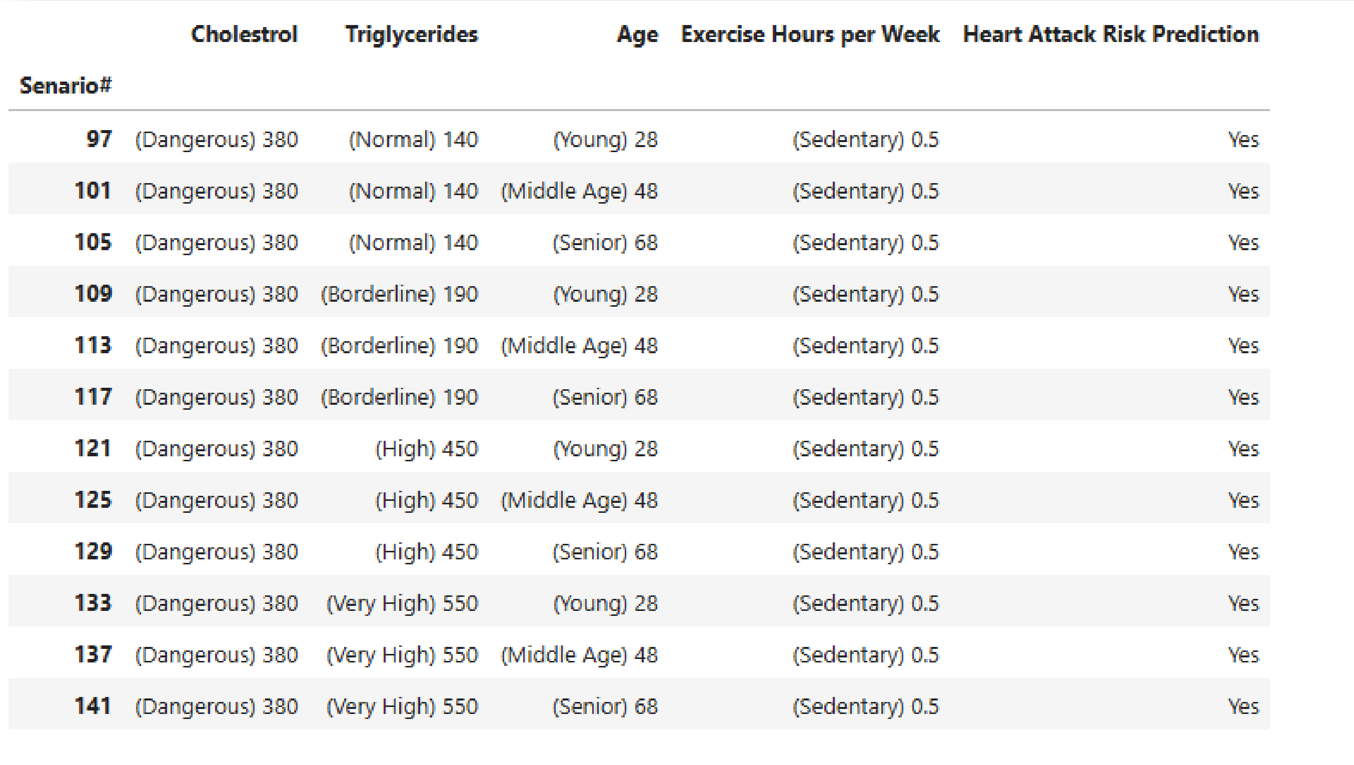


Feature Importance Analysis

Scenarios Tested

- Cholesterol Levels:
  - Normal (180)
  - At Risk (230)
  - Dangerous (380)
- Triglycerides Levels:
  - Normal (140)
  - Borderline (190)
  - High (450)
  - Very High (550)

- Age Groups:
  - Young (28)
  - Middle Age (48)
  - Senior (68)
- Activity Levels:
  - Sedentary (0.5 hrs/week)
  - Lightly Active (2 hrs/week)
  - Moderately Active (4 hrs/week)
  - Very Active (6 hrs/week)



	Cholestrol	Triglycerides	Age	Exercise Hours per Week	Heart Attack Risk Prediction
Senario#					
97	(Dangerous) 380	(Normal) 140	(Young) 28	(Sedentary) 0.5	Yes
101	(Dangerous) 380	(Normal) 140	(Middle Age) 48	(Sedentary) 0.5	Yes
105	(Dangerous) 380	(Normal) 140	(Senior) 68	(Sedentary) 0.5	Yes
109	(Dangerous) 380	(Borderline) 190	(Young) 28	(Sedentary) 0.5	Yes
113	(Dangerous) 380	(Borderline) 190	(Middle Age) 48	(Sedentary) 0.5	Yes
117	(Dangerous) 380	(Borderline) 190	(Senior) 68	(Sedentary) 0.5	Yes
121	(Dangerous) 380	(High) 450	(Young) 28	(Sedentary) 0.5	Yes
125	(Dangerous) 380	(High) 450	(Middle Age) 48	(Sedentary) 0.5	Yes
129	(Dangerous) 380	(High) 450	(Senior) 68	(Sedentary) 0.5	Yes
133	(Dangerous) 380	(Very High) 550	(Young) 28	(Sedentary) 0.5	Yes
137	(Dangerous) 380	(Very High) 550	(Middle Age) 48	(Sedentary) 0.5	Yes
141	(Dangerous) 380	(Very High) 550	(Senior) 68	(Sedentary) 0.5	Yes

Scenarios with Heart Attack Risk Predicted as Yes

Key Insights from the Simulations

High Cholesterol & Triglycerides increased risk significantly.

Sedentary Lifestyle (0.5 hours/week exercise) was a major risk factor.

Being at least lightly active (2+ hrs/week) reduced risk even with high cholesterol/triglycerides.

Summary & Next Steps

Current Accuracy: 62% (Decision Tree model).

Challenges: Low class separation (ROC AUC Score ~0.50).

Potential improvements:

- Use ensemble learning (XGBoost, stacking).
- Expand dataset with additional health factors.
- Optimize feature selection and hyperparameters.