4/15/2025

# Guided Capstone Two Project Report
# Heart Attack Risk Prediction

Nilesh Suresh

## Contents

# 1. Context

The Heart Attack Risk Prediction Dataset helps researchers and doctors better understand heart health and its risk factors. Heart attacks are a serious global health problem, and it's important to identify what leads to them and how to prevent them.

This dataset includes information like age, cholesterol levels, blood pressure, smoking habits, exercise routines, diet choices, and more. By analyzing these details with **predictive analytics and machine learning**, experts can find patterns that help predict heart attack risk. This knowledge allows them to develop **better prevention strategies** and improve heart health worldwide.

In short, the dataset is a **valuable tool** for studying heart disease, making early predictions, and helping people live healthier lives.

Heart Attack Prediction Dataset

# 2. Problem Statement

Leverage **predictive analytics and machine learning** techniques to analyze the **Heart Attack Risk Prediction Dataset**, identifying patterns and correlations among various features that contribute to heart attack risk. By developing an accurate and interpretable predictive model, healthcare professionals and researchers can gain deeper insights into risk factors and formulate **proactive strategies for prevention and management**. The goal is to contribute to **improved cardiovascular health**, enabling timely interventions and fostering a data-driven approach to heart disease prevention.

# 3. Data Wrangling

Data Wrangling Report

The dataset was first loaded and set with "Patient ID" as the index, consisting of 8,763 rows and 25 columns. A thorough exploration confirmed that there were no missing values, and summary statistics of numeric features were reviewed.

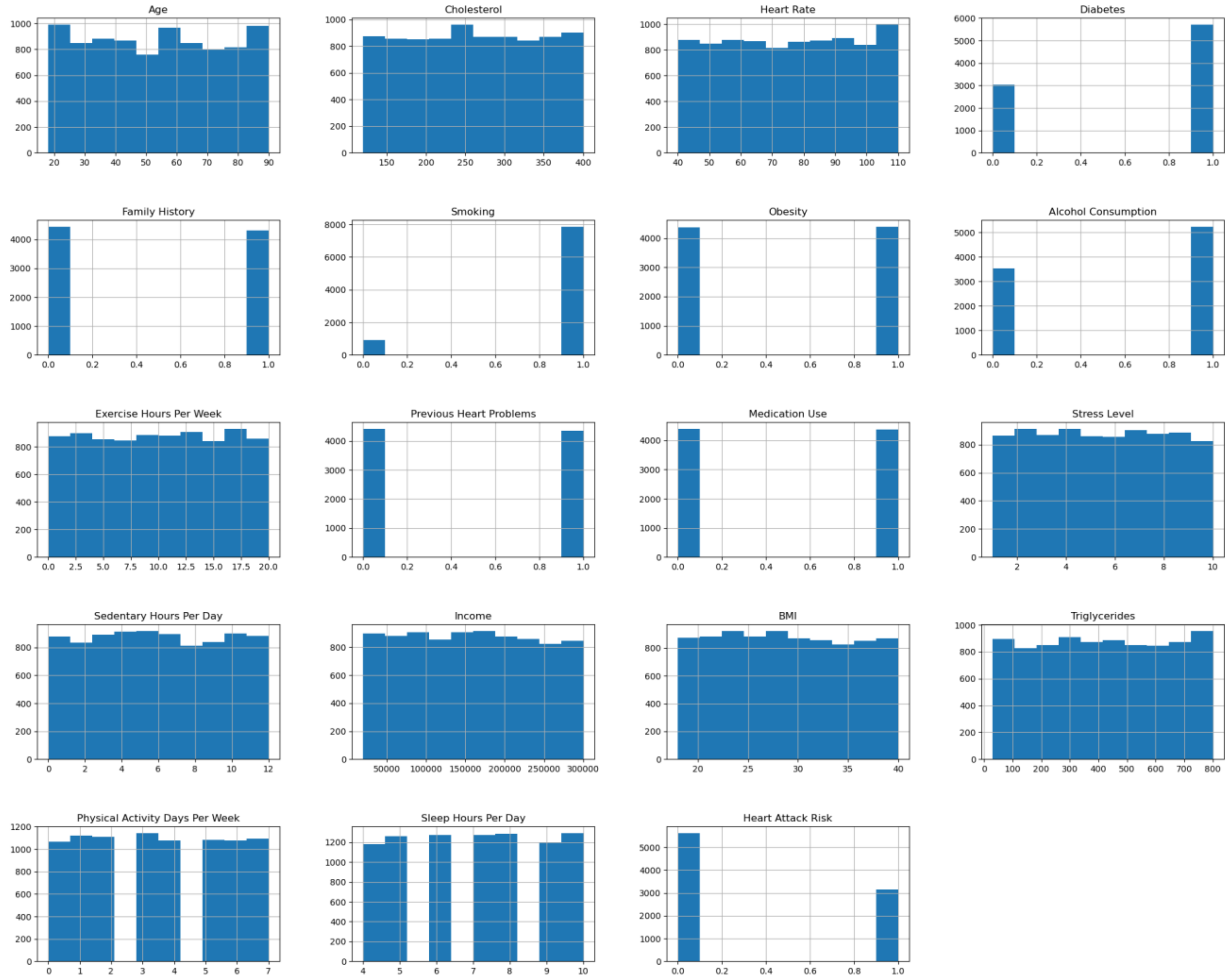
## 3.1 Cleaning steps included

1. Rounding values for Exercise Hours Per Week, Sedentary Hours Per Day, and BMI to ensure consistency.

2. Blood Pressure, initially a single string column, was split into Systolic and Diastolic, then converted to integers for better analysis.

3. Categorical features such as Sex, Diet, Country, Continent, and Hemisphere were validated and found suitable for further steps.

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



**"Heart Attack Risk"** is the key target feature for predictive modeling. This step ensures the dataset is clean, structured, and ready for deeper analysis and predictive modeling.

# 4. Exploratory Data Analysis

EDA Report

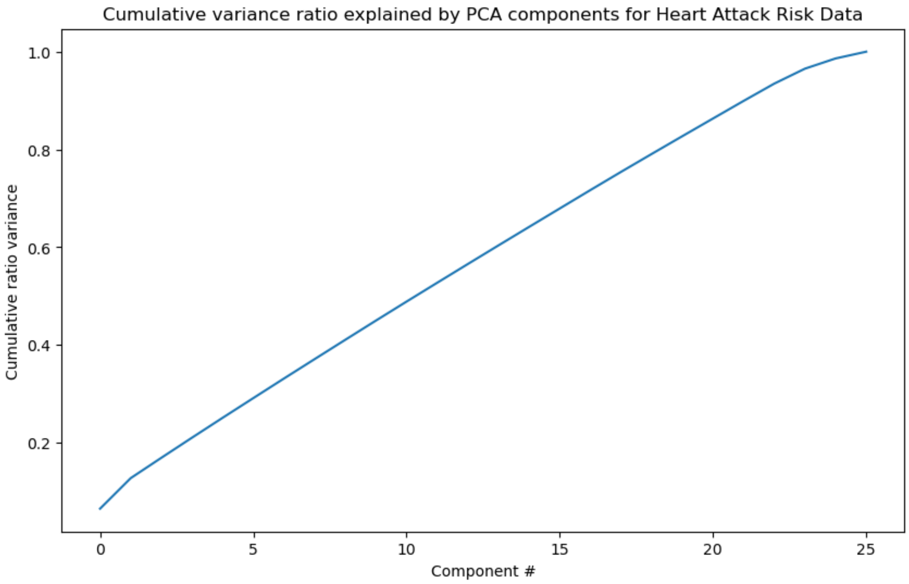
The Exploratory Data Analysis (EDA) step provides deeper insights into the dataset, helping to identify trends, anomalies, and relationships between variables before building predictive models. This process ensures a solid foundation for further analysis.

## 4.1 Data Cleaning & Transformation

- Label Encoding was applied to categorical variables (Sex, Diet, Country, Continent, Hemisphere) to enable visualization.
- Blood Pressure values (stored as strings) were split into Systolic and Diastolic, then converted into numerical values.
- Encoded categorical variables replaced the original columns for streamlined analysis.

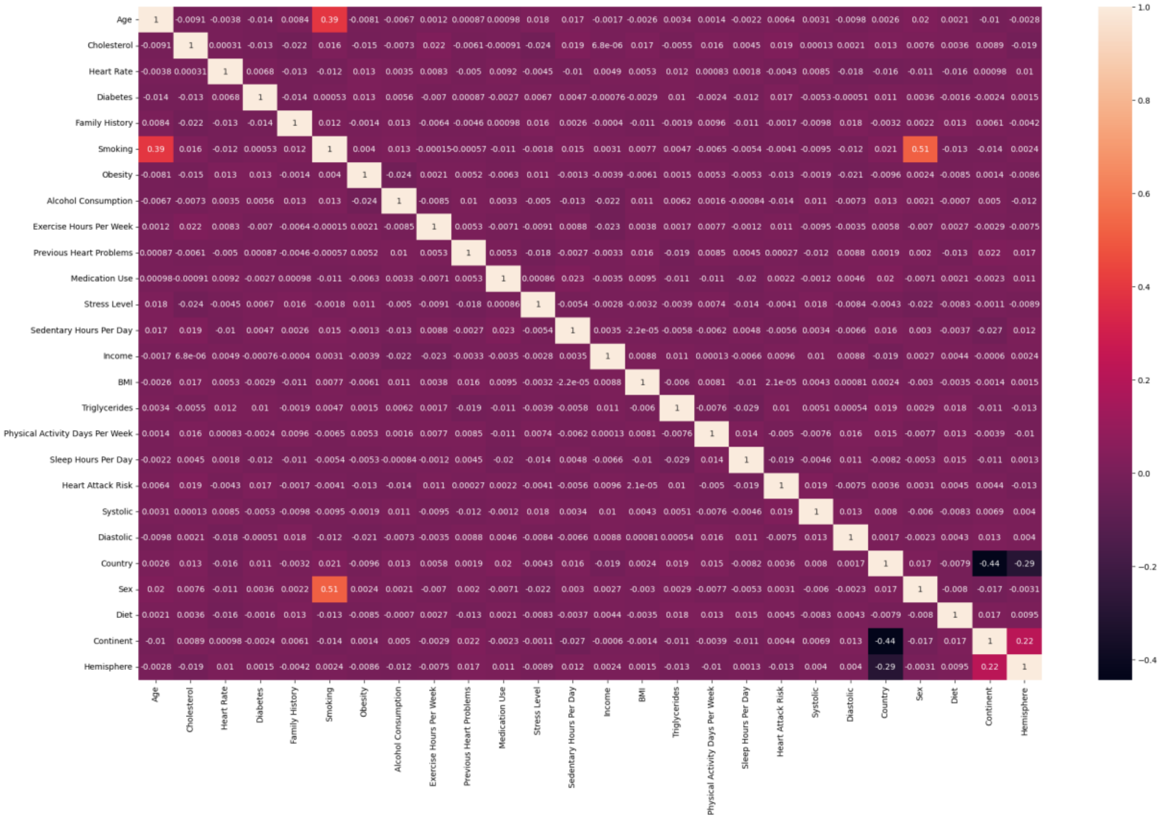
## 4.2 Visualization & Feature Analysis

1. 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.



2. 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



3. Feature Mapping for Better Interpretability
    - Mapped Cholesterol values into "Healthy," "At Risk," and "Dangerous" categories.
    - Categorized Systolic Blood Pressure into risk levels (Healthy, Elevated, Hypertension Stages 1 & 2, Crisis).

## 4.3 Key Findings

- High correlation between "Heart Attack Risk" and:
    - Cholesterol, Diabetes, Blood Pressure, Age, Previous Heart Problems, and Medication Use.
- Older patients with high cholesterol and blood pressure have greater heart attack risk.
- Alcohol consumption does not show a strong link to heart attack risk.
- Smoking does not emerge as a major predictor of heart attack risk compared to other factors.

EDA successfully identified **key features influencing heart attack risk**, refined the dataset for predictive modeling, and highlighted important trends. These insights will guide the next steps in **building machine learning models** to enhance heart disease risk assessment.

# 5. Pre-Processing and Training Data

Preprocessing and Training Report

This step involves preparing the dataset for machine learning by encoding categorical features, ensuring all data is numeric, and splitting it into training and testing subsets for model development.

## 5.1 Dummy Feature Creation (Encoding Categorical Variables)

- Cholesterol Level and Blood Pressure, originally categorical, were converted to numeric dummy features using one-hot encoding (pd.get_dummies()).
- This ensures machine learning models can process categorical data effectively.
- A check was performed to confirm all columns are now numerical.

## 5.2 Train/Test Split

- The dataset split 80/20 into training (80%) and testing (20%) subsets.
- Feature and Target Separation:
    - X (Features) → Contains all features except "Heart Attack Risk".
    - y (Target Variable) → "Heart Attack Risk", which the model will predict.
- train_test_split() was used with random_state=42 to ensure reproducibility.

This step successfully transformed categorical variables into numerical features, ensuring compatibility with machine learning algorithms. The train/test split establishes a structured approach for model training and validation. The dataset is now ready for model selection and evaluation in the next phase.

# 6. Modeling

Modeling Report

This report details the modeling approaches used to predict heart attack risk. Five classification models—Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Decision Tree, and Gradient Boosting—were trained and evaluated using key metrics such as accuracy, confusion matrix, classification report, and ROC AUC score.

## 6.1 Models & Evaluation

### 6.1.1 Logistic Regression

- Best hyperparameter: C = 0.2113
- Accuracy: 64.1%
- Confusion Matrix:
  - Correctly predicted: 4484 No-Risk, 13 Risk cases.
  - Misclassified: 2500 No-Risk as Risk, 14 Risk as No-Risk.
- ROC AUC Score: 0.4978 (poor class separation).

### 6.1.2 Random Forest Classifier

- Best hyperparameters: Criterion: gini, Max Depth: 3, Min Samples Split: 2, N Estimators: 200
- Accuracy: 64.1%
- Confusion Matrix:
  - Correctly predicted: 4497 No-Risk, 0 Risk cases.
  - Misclassified: 2513 No-Risk as Risk, 1 Risk as No-Risk.
- ROC AUC Score: 0.5079 (slightly better but still weak).

### 6.1.3 KNeighbors Classifier

- Best hyperparameter: n_neighbors = 47
- Accuracy: 63.8%
- Confusion Matrix:

    - Correctly predicted: 4452 No-Risk, 20 Risk cases.
    - Misclassified: 2493 No-Risk as Risk, 46 Risk as No-Risk.
- ROC AUC Score: 0.5070 (similar to Random Forest).

### 6.1.4 Decision Tree Classifier

- Best hyperparameters: Criterion: entropy, Max Depth: 3, Min Samples Split: 2
- Accuracy: 61.9%
- Confusion Matrix:
    - Correctly predicted: 4135 No-Risk, 207 Risk cases.
    - Misclassified: 2306 No-Risk as Risk, 363 Risk as No-Risk.
- ROC AUC Score: 0.5027 (weak class separation).

### 6.1.5 Gradient Boosting Classifier

- Best hyperparameters: Max Depth: 3, Min Samples Split: 2, Learning Rate: 0.1
- Accuracy: 63.0%
- Confusion Matrix:
    - Correctly predicted: 4322 No-Risk, 97 Risk cases.
    - Misclassified: 2416 No-Risk as Risk, 176 Risk as No-Risk.
- ROC AUC Score: 0.5134 (slightly better, but weak).

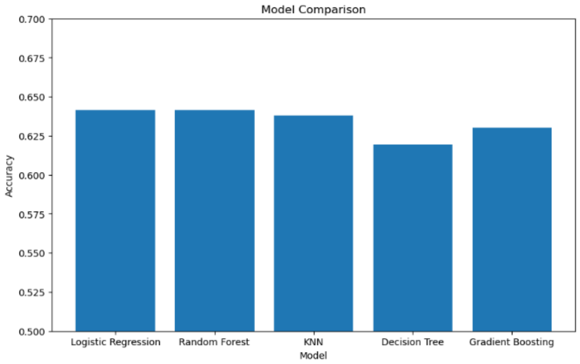
## 6.2 Model Comparisons & Insights

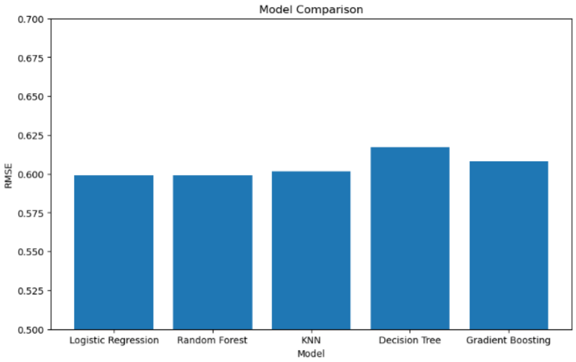
### 6.2.1 Comparison of Key Metrics

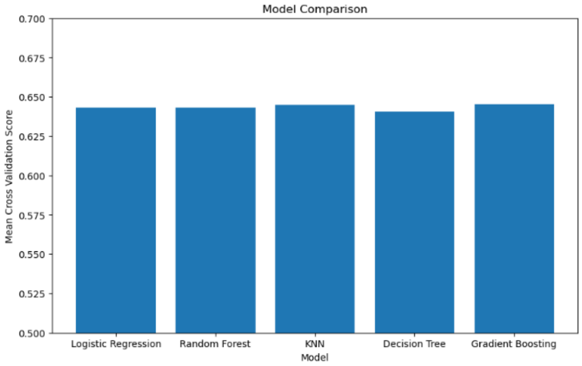
- Accuracy: Logistic Regression & Random Forest performed best (64.1%), followed by Gradient Boosting & KNN (63.0%).
- ROC AUC Score: Gradient Boosting had the highest score at 0.5134, but overall models struggle with class separation.
- Confusion Matrix Analysis: Decision Tree and Gradient Boosting correctly predicted more heart attack risk cases but at the cost of higher false positives.

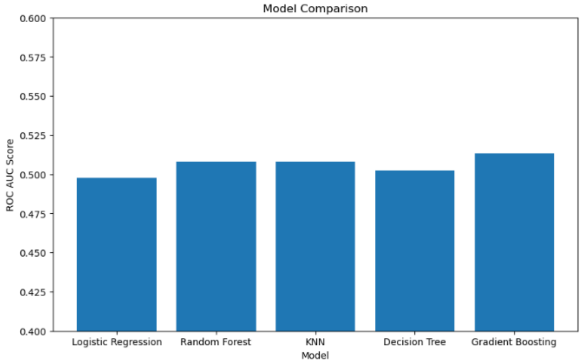
## 6.2.2 Visualization Summary

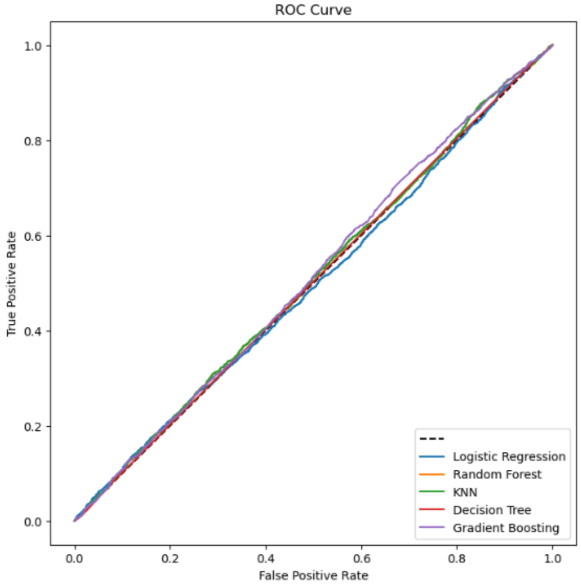
Each model's accuracy, RMSE, cross-validation score, and ROC AUC score were visualized. The trends suggest all models need improvement to differentiate between positive and negative cases effectively.











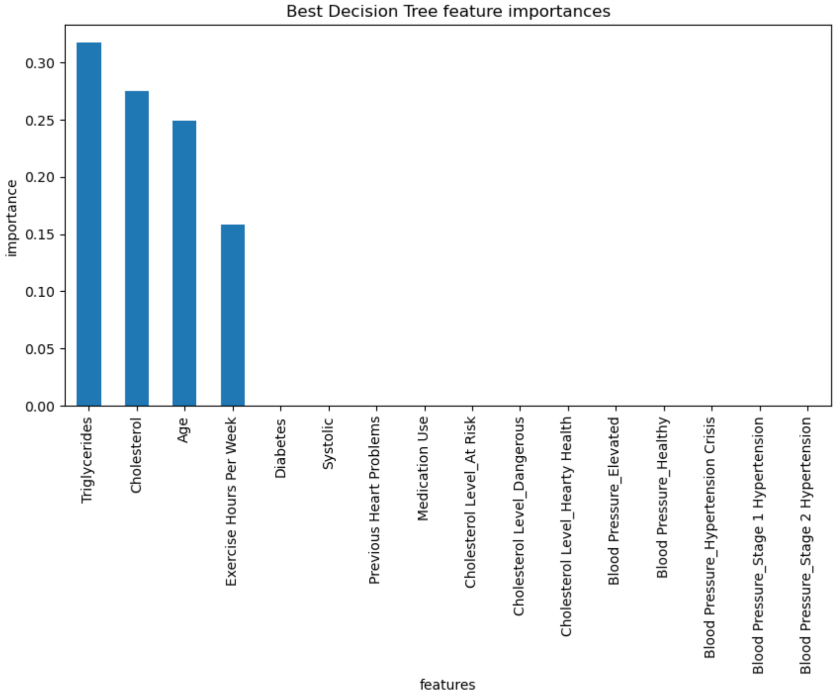
## 6.3 Final Model Selection

After evaluating multiple models based on accuracy, classification reports, confusion matrices, and ROC AUC scores, **Decision Tree Classifier** was selected as the final model for heart attack risk prediction. This decision was based on:

- **Performance Comparison:** Gradient Boosting and Decision Tree performed best overall.
- **Prediction Ability:** Decision Tree correctly predicted **2.95% of heart attack risk cases**, making it the strongest model for identifying positive instances.

### 6.3.1 Feature Importance Analysis

Using the trained Decision Tree model, feature importance was plotted.



This reveals the primary indicators for heart attack risk prediction:

- **Triglycerides**
- **Cholesterol**
- **Age**
- **Exercise Hours Per Week**

These variables were most influential in determining risk levels, demonstrating their critical impact on heart health.

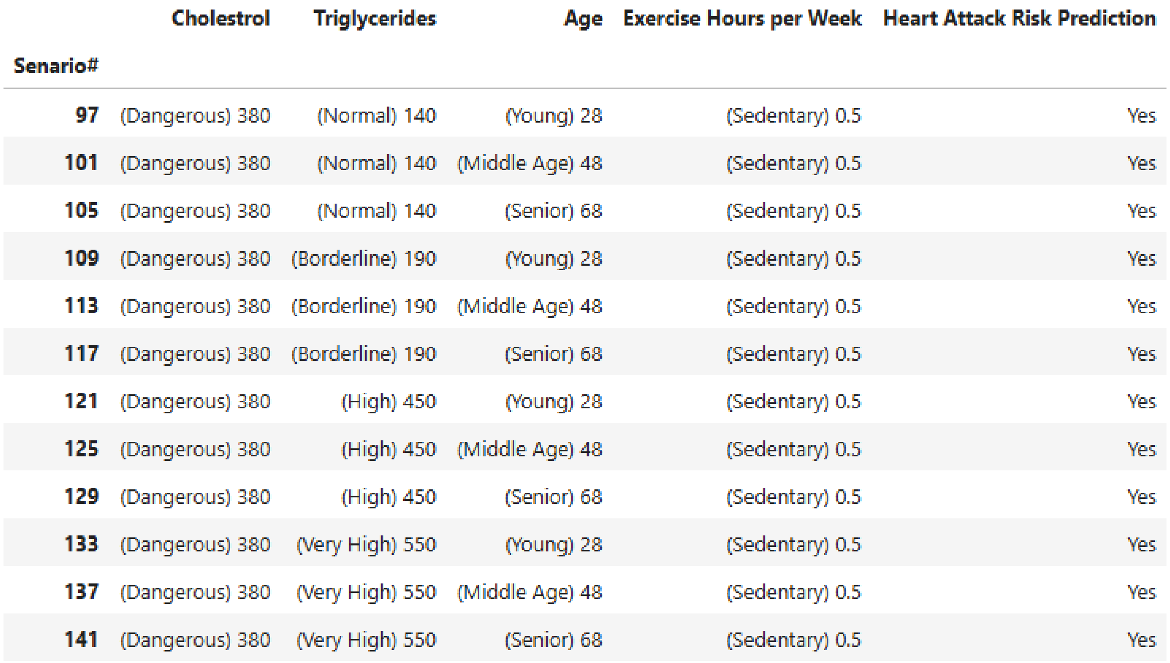
## 6.3.2 Modeling Scenarios

A function (predictHeartAttackRisk) was created to **simulate heart attack risk predictions** under various scenarios based on key health metrics.

**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)

**Scenarios with Heart Attack Risk Predicted as Yes:**



| | 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 |

**Key Insights from the Simulations**

1. **High Cholesterol & Triglycerides** increased risk significantly.
2. **Sedentary Lifestyle** (0.5 hours/week exercise) was a major risk factor.
3. **Being at least lightly active (2+ hrs/week)** reduced risk even with high cholesterol/triglycerides.

## 7. Summary & Next Steps

- The **Decision Tree model achieved 62% accuracy** but still struggled with clear class separation (ROC AUC Score: 0.5027).
- While useful for identifying some risk factors, the model lacks strong predictive power (**recommended AUC > 0.8**).
- **Feature expansion** and advanced techniques like **ensemble learning** (e.g., XGBoost, stacking models) may improve results.
- **Additional data collection** (dietary habits, genetics, lifestyle specifics) could enhance predictive strength.