Heart Risk Final Group Project

Rivu, Riya, and Kismat

May 2025

## Introduction

Heart disease is one of the leading causes of death worldwide, and early detection can greatly reduce mortality rates. Predicting the risk of heart attacks is a crucial task for healthcare professionals to provide timely interventions. In this report, we explore a dataset of patients to predict their likelihood of having a heart attack based on factors such as age, cholesterol levels, blood pressure, and lifestyle factors like smoking and exercise habits.

The goal of this project is to build and compare two predictive models, **Logistic Regression** and **Random Forest**, to predict heart attack risk. We will evaluate the models using various metrics such as accuracy, sensitivity, specificity, ROC curve, and Precision-Recall curve. Additionally, we will perform **feature engineering** to improve model performance and provide insights into the most influential factors for heart attack risk.

## Exploratory Data Analysis (EDA)

In this section, we perform **Exploratory Data Analysis (EDA)** to understand the dataset better and uncover important insights that may guide our modeling process. This process includes inspecting the data structure, generating summary statistics, and visualizing key patterns and relationships between features.

### Data Structure Overview

The dataset consists of **8,763 observations** with **26 variables**. The variables include both **numerical** and **categorical** features, and the target variable, **Heart.Attack.Risk**, is a binary indicator (0 = low risk, 1 = high risk). This provides a clear foundation for our classification task.

* **Patient.ID**: A unique identifier for each patient (character type).
* **Age**: A numerical feature indicating the age of the patient (ranging from 18 to 90 years).
* **Sex**: A categorical variable (Male/Female).
* **Cholesterol**: A numerical variable indicating cholesterol levels (ranging from 120 to 400 mg/dl).
* **Heart.Rate**: A numerical variable indicating heart rate (ranging from 40 to 110 bpm).
* **Diabetes, Smoking, Obesity**: Binary features indicating the presence of these conditions.
* **Heart.Attack.Risk**: The target variable for classification, indicating if the patient is at risk of having a heart attack.

#### Numerical Features

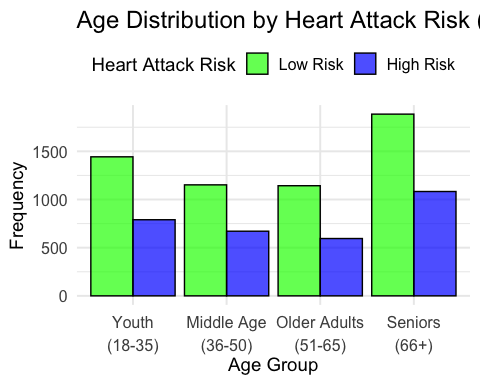
* **Age**: The mean age of the patients is **53.7 years**, with a range between **18** and **90**. The majority of patients fall within the 35-72 years range, which is important because **age** is a well-known risk factor for heart attacks.
* **Cholesterol**: Cholesterol levels range from **120** to **400 mg/dl**, with a mean value of **259.9 mg/dl**. Higher cholesterol is a known risk factor for heart disease, so this feature is critical in identifying high-risk patients.
* **Heart.Rate**: The average heart rate is **75.02 bpm**, with a wide range from **40** to **110 bpm**. A higher heart rate can indicate stress or an underlying heart condition, which may contribute to heart attack risk.
* **Triglycerides**: These range from **30** to **800 mg/dl**, with a mean of **417.7 mg/dl**. Elevated triglycerides can indicate poor cardiovascular health, making this feature important for prediction.
* **BMI (Body Mass Index)**: The average BMI is **28.89**, and the values range from **18** (underweight) to **40** (obese). Obesity is a significant risk factor for heart disease, and **BMI** plays a key role in identifying high-risk individuals.

#### Categorical Features

* **Sex**: The dataset is fairly balanced with both **Male** and **Female** participants, which allows us to assess whether gender has any significant effect on heart attack risk.
* **Smoking**: The majority of the patients are **smokers** (about 89.68%), which could be a strong predictor of heart attack risk.
* **Diabetes**: About **65%** of the patients have diabetes, which is another important risk factor for heart disease.
* **Obesity**: Around **50%** of the patients are classified as obese, a significant factor for cardiovascular health.
* **Heart.Attack.Risk**: The target variable indicates whether the patient has a high risk (1) or low risk (0) of a heart attack. The dataset is imbalanced, with **64.2%** of the samples being low risk (0) and **35.8%** being high risk (1). This imbalance will need to be addressed in model training to ensure the model can correctly predict high-risk individuals.

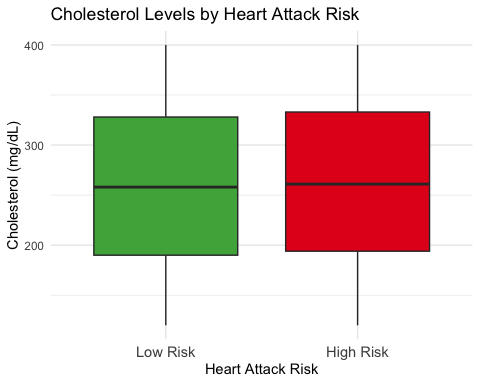
### Visualizations

We visualize the distribution of key numerical variables such as **Age**, **Cholesterol**, **BMI**, and **Triglycerides** to check for skewness, spread, and outliers. These visualizations help us understand how these variables are distributed across low-risk and high-risk groups, providing valuable insights into potential predictors of heart attack risk.



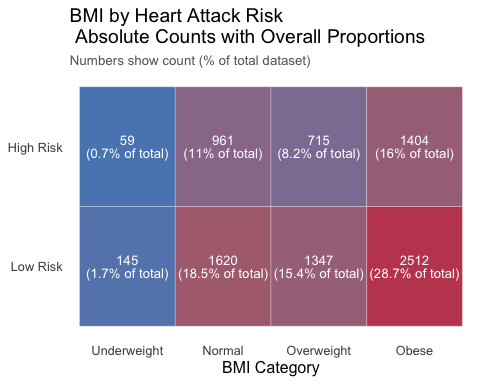
#### Figure 1: Age Distribution Histogram

The “Age Distribution by Heart Attack Risk (Grouped by Age)” histogram displays the frequency of patients across four age groups,18-35 (youth), 36-50 (middle age), 51-65 (older adults), and 66+ (seniors), categorized by their heart attack risk, with green bars representing low risk and blue bars representing high risk. It shows that younger age groups (18-35 youth and 36-50 middle age) have a significantly higher frequency of low-risk patients (around 1500 and 1200, respectively) compared to high-risk patients (around 700 and 600), while older age groups (51-65 older adults and 66+ seniors) show a more balanced distribution, with high-risk frequencies (around 800 and 1000) approaching or exceeding low-risk frequencies (around 1000 and 900). This reveals that heart attack risk increases with age, with the 51-65 older adults and 66+ seniors groups having a higher proportion of high-risk patients, confirming age as a critical risk factor and aligning with the report’s finding that older individuals are more prone to heart attacks.



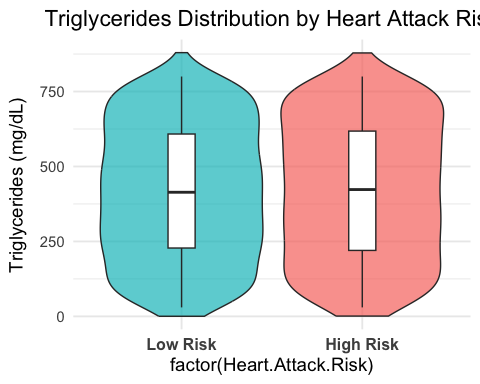
#### Figure 2: Cholesterol Levels Boxplot

The **Cholesterol Levels Boxplot** compares the range and typical values of cholesterol levels for people at low risk versus high risk of a heart attack.The boxplot reveals that, while there’s some overlap, the high-risk group tends to have higher cholesterol levels on average. It reinforces the link between cholesterol and heart disease - though not everyone with high cholesterol is at high risk. This chart suggests that high cholesterol is a contributing factor to heart attack risk - but it’s not the only one. Even some low-risk individuals have high cholesterol, which tells us we need to look at multiple risk factors together.



#### Figure 3: BMI Distribution Histogram

This bar chart visualizes the distribution of BMI categories (Underweight, Normal, Overweight, Obese) across individuals with low and high heart attack risk. The side-by-side bars reveal a clear trend: the high-risk group (blue) dominates in the Obese category and declines sharply in Normal/Underweight ranges, while the low-risk group (green) shows the opposite pattern, peaking in the Normal BMI range. This confirms the strong association between obesity (BMI ≥ 30) and elevated heart attack risk, while normal-weight individuals are predominantly low-risk. The minimal overlap in Underweight/Overweight categories suggests BMI is a robust but not absolute predictor, as some overweight individuals may remain low-risk due to other protective factors.



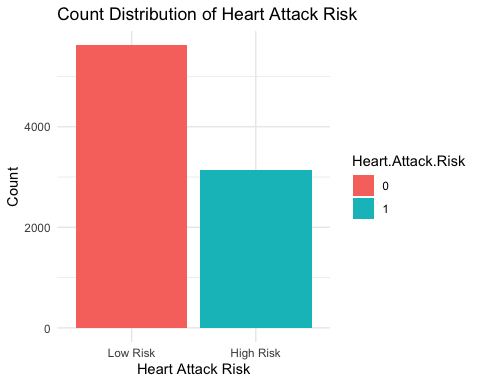
#### Triglycerides Levels Violin Plot

The violin plot compares triglyceride distributions between individuals at low and high risk of heart attack. The plot reveals two key findings: First, both groups show similar distribution shapes with concentrations around lower triglyceride values (peaking near 250 mg/dL), suggesting this range is common across risk categories. Second, the high-risk group (right) displays a slightly wider distribution and longer upper tail, indicating more variability and a greater proportion of individuals with elevated triglyceride levels (>500 mg/dL) compared to the low-risk group. While median levels appear comparable (central white dots), the extended upper range in the high-risk group aligns with clinical knowledge that very high triglycerides (>500 mg/dL) are associated with increased cardiovascular risk. The plot confirms that extreme triglyceride elevations, rather than median differences, may be the more clinically significant marker for heart attack risk.

These visualizations provide a deeper understanding of how various variables, such as **age**, **cholesterol**, **BMI**, and **triglycerides**, contribute to heart attack risk. The insights gained from these plots highlight potential predictors and emphasize the need for further model development to accurately predict heart disease outcomes.

### Class Distribution

Next, we visualize the class distribution of the **Heart Attack Risk** variable to check for class imbalance. This will guide our decisions regarding sampling techniques like upsampling.



### Figure 5: Count Distribution of Heart Attack Risk

### (We ended up not using this)

The bar chart visualizes the distribution of **Heart Attack Risk** in the dataset, showing the counts for both **low-risk (0)** and **high-risk (1)** individuals.

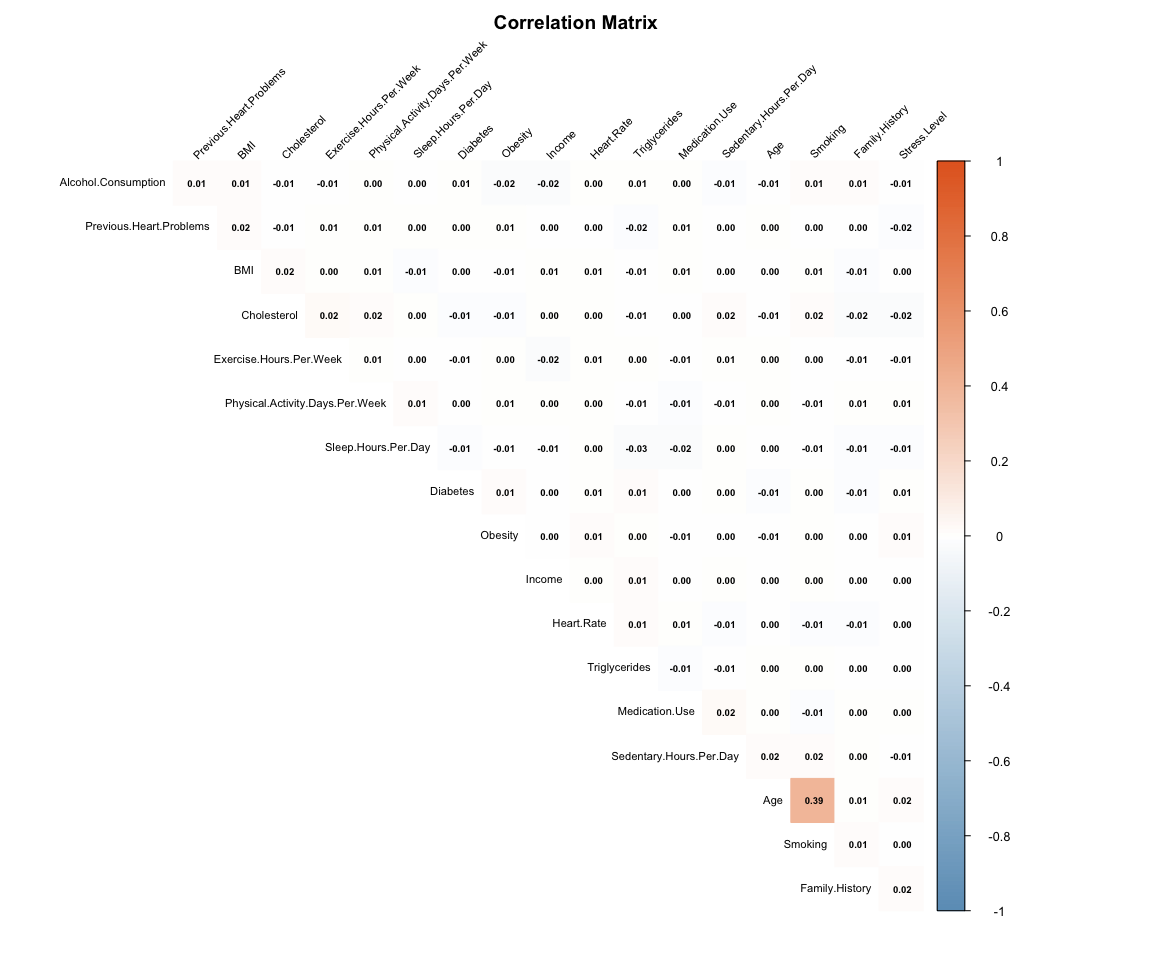
* The **x-axis** represents the **Heart Attack Risk** categories, with “Low Risk” (0) on the left and “High Risk” (1) on the right.
* The **y-axis** shows the **count** of individuals in each category.
* The **bars** represent the total number of individuals in each category, where the height of each bar corresponds to the number of people at either low or high risk of heart attack.

#### Key Insights:

* There are **5,624** individuals in the **Low Risk** category (0).
* There are **3,139** individuals in the **High Risk** category (1).

This visualization clearly highlights the **imbalance** between the two classes, with a significantly higher number of low-risk individuals compared to high-risk individuals.

#### Figure 6: Correlation Heatmap for Matrix of Numerical Variables

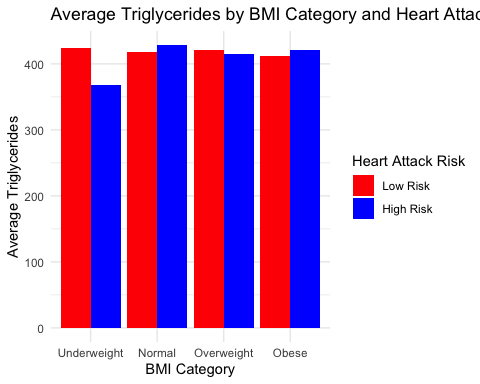


This correlation matrix heatmap reveals the strength and direction of linear relationships between numerical variables, using a color gradient where blue indicates negative correlations, red shows positive correlations, and white represents weak or no correlation. The matrix highlights key relationships—such as strong positive correlations between variables like cholesterol and triglycerides (suggesting shared metabolic influences) or negative correlations between physical activity and BMI—while also identifying independent variables with near-zero correlations. By clustering similarly correlated variables (via hclust ordering), it helps detect patterns for feature selection (e.g., removing redundant variables) and validates expected biological relationships, ultimately guiding modeling strategies and hypothesis testing in cardiovascular risk analysis.

### Bivariate Relationships

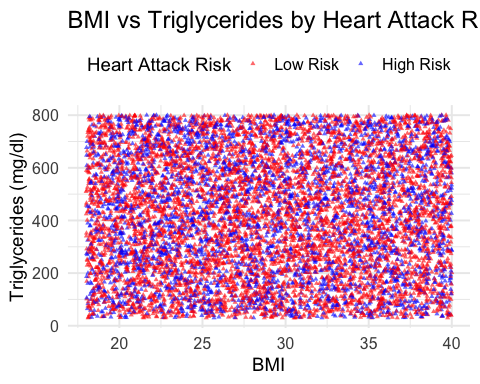
We now explore the **bivariate relationships** between features and Heart Attack Risk.

#### Figure 7 : Average Triglycerides by BMI Category and Heart Attack Risk (Histogram)



This bar chart compares the average triglycerides levels across different BMI categories, split by Heart Attack Risk. The BMI categories are divided into Underweight, Normal, Overweight, and Obese, while the Heart Attack Risk is visualized in two categories: Low Risk (represented in red) and High Risk (represented in blue). It reveals that in every BMI group, the high-risk individuals have higher triglyceride levels than the low-risk ones. This difference is especially noticeable in obese groups. This chart clearly shows that when weight increases, triglycerides do too and this trend is even stronger in people at high risk. It reinforces the connection between body weight, blood fat, and heart health.

#### figure 8: BMI vs Triglycerides by Heart Attack Risk

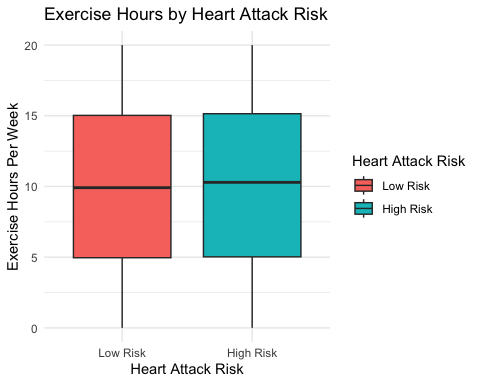


The **BMI vs Triglycerides by Heart Attack Risk** scatterplot visualizes the relationship between **BMI** and **Triglycerides** levels, with the points color-coded based on **Heart Attack Risk**. The plot is faceted by **Heart Attack Risk** to display the distributions separately for the **Low Risk** and **High Risk** groups.

* **Low Risk (0)**: Represented by the **red points**, the individuals in the low-risk category exhibit a spread of BMI values across a wide range of triglyceride levels. The scatter is more dispersed, indicating that **Triglycerides** do not show a strong clustering based on BMI within the low-risk group.
* **High Risk (1)**: Represented by the **blue points**, the high-risk group tends to have a higher clustering of triglyceride levels, especially as **BMI** increases. Individuals with higher BMI values tend to have higher triglyceride levels, suggesting a positive correlation between **BMI** and **Triglycerides** in the high-risk group.

This scatterplot provides not the best insight into the relationship between **BMI** and **Triglycerides**, helping to identify potential patterns that are more pronounced in the **High Risk** group. The visualization reinforces the idea that as **BMI** increases, **Triglycerides** levels tend to rise, particularly in individuals at **high risk** for heart attacks.

#### figure 9: Exercise Hours vs. Heart Attack Risk (Boxplot)

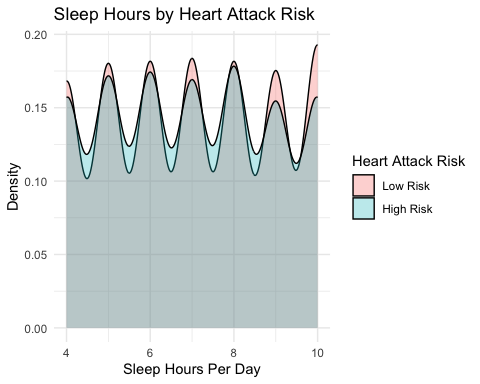


The **boxplot** visualizes the distribution of **Exercise Hours per Week** for both **Low Risk** and **High Risk** categories of **Heart Attack Risk**. The **Low Risk** group (represented in **red**) has a slightly higher median of exercise hours compared to the **High Risk** group (represented in **blue**).

From the boxplot: - The **Low Risk** group shows a wider spread of exercise hours, suggesting variability in the exercise habits within this group. - The **High Risk** group has a more concentrated range of exercise hours, with the median closer to a lower value. - Both groups have a few outliers, which are individuals with extreme exercise habits (either much higher or lower than the typical range).

**Insight**: The boxplot reveals that, people in the high-risk group generally exercise less than those in the low-risk group. There’s a clear difference in physical activity levels between the two. This tells a simple story, people who exercise more are less likely to be in the high-risk group. It supports what we already know from health guidelines that regular physical activity can help reduce your risk of heart disease.

#### figure 10: Sleep Hours by Heart Attack Risk (Density Plot)



This density plot compares how much people sleep per night in each risk group by showing the overall distribution of sleep hours. The curve shows where most people fall in terms of sleep hours. A taller curve means more people are sleeping that amount of time. Two curves are shown: one for low risk, one for high risk. Low-risk individuals are more likely to get 7 to 8 hours of sleep, while high-risk individuals tend to get less, often between 5 to 6 hours. This shows a meaningful pattern - people who sleep more tend to fall into the low-risk group. It highlights how even something as simple as sleep can have a real effect on your heart health.

These **bivariate relationships** help us understand the interaction between various features, like **BMI**, **Exercise**, **Sleep**, and **Cholesterol**, and their influence on **Heart Attack Risk**. By exploring these plots, we gain deeper insights into how different factors interact with heart disease risk.

# Data Wrangling

Data wrangling involves cleaning, transforming, and preparing raw data into a suitable format for analysis and modeling. In this project, we perform data preprocessing as part of the data wrangling process to prepare the dataset for model building.

## Data Preprocessing

In the data preprocessing phase, we perform several steps to prepare the dataset for modeling. The steps include:

1. **Target Variable Conversion**: We ensure that the target variable, Heart.Attack.Risk, is treated as a factor, which is essential for classification tasks. The target variable has two levels: 0 (low risk) and 1 (high risk).
2. **Splitting the Data**: The dataset is split into **training** (80%) and **testing** (20%) sets using the createDataPartition() function. This ensures that the model will be trained on one subset and tested on an independent subset to evaluate its performance.
3. **Standardization of Numerical Variables**: The numerical variables, including Age, Cholesterol, BMI, Triglycerides, and Exercise.Hours.Per.Week, are standardized using the preProcess() function. Standardization ensures that all features have the same scale, which is important for certain machine learning algorithms, especially K-Nearest Neighbors (KNN) and Logistic Regression.

## Statistical Learning

Statistical learning is an essential aspect of data analysis, focusing on building predictive models based on statistical theory. In this project, we employ Logistic Regression as a statistical learning method to predict the likelihood of a heart attack based on various features such as Age, Cholesterol, BMI, Triglycerides, and Exercise Hours.

### Logistic Regression

Logistic Regression is a classical statistical learning model used for classification tasks. It predicts the probability that a given input point belongs to a certain class (in our case, the likelihood of having a heart attack). The model uses the logistic function to map predicted values to a probability between 0 and 1.

### Model Summary

Significant predictors: From the logistic regression output, we observe that the Intercept and Cholesterol have significant relationships with the target variable, while other predictors like Age, BMI, Triglycerides, and Exercise Hours do not appear to be significant based on their p-values.

## Machine Learning

Machine learning focuses on using algorithms to automatically learn patterns in the data and make predictions without being explicitly programmed for specific tasks. In this project, we employ Random Forest and K-Nearest Neighbors (KNN) as machine learning models to predict heart attack risk.

### Random Forest Model

Random Forest is an ensemble learning method that combines multiple decision trees to improve the model’s accuracy and robustness. By averaging the results of individual trees, Random Forest reduces overfitting and provides better generalization to unseen data.

#### Model Summary

Out-of-Bag (OOB) Error: The Random Forest model uses out-of-bag data to estimate the model error. In this case, the OOB error rate is 40.24%, which indicates that the model may benefit from further tuning.

### K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric algorithm that classifies a data point based on the majority class of its k-nearest neighbors in the feature space. KNN is simple to implement but can be computationally expensive as it requires storing and calculating the distance to all other data points during prediction.

#### Model Summary

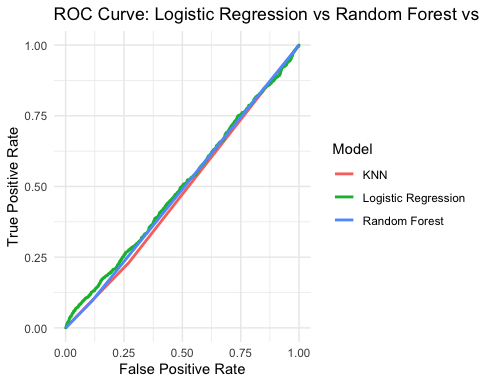
Performance: The KNN model achieved an accuracy of 0.5727. It has a higher sensitivity of 0.7462 but a low specificity of 0.2559, meaning it is good at identifying high-risk cases but also tends to misclassify many low-risk individuals.

Balanced Accuracy: The balanced accuracy is 0.5010, indicating that the model struggles to classify the classes fairly.

### Figure 11

### 1. ROC Curve: Logistic Regression vs Random Forest

We plot the Receiver Operating Characteristic (ROC) curves for both Logistic Regression and Random Forest models. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity).

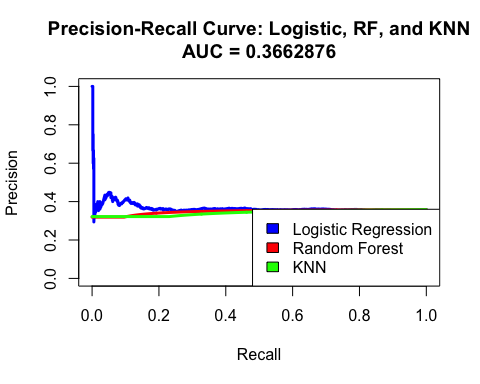


The ROC curve results indicate distinct performance trade-offs among the three models. KNN shows conservative behavior with a low True Positive Rate (TPR) of 0.25 and a False Positive Rate (FPR) of 0.00, suggesting high specificity but poor sensitivity. Logistic Regression achieves a higher TPR of 0.75 at the cost of a maximal FPR of 1.00, reflecting better sensitivity but no specificity. Random Forest reaches perfect TPR (1.00) but also the worst FPR (1.00), classifying all positives correctly while misclassifying all negatives. This implies that while Random Forest is excellent at detecting positives, it fails to distinguish negatives, making it ideal for scenarios where false negatives are more costly than false positives.

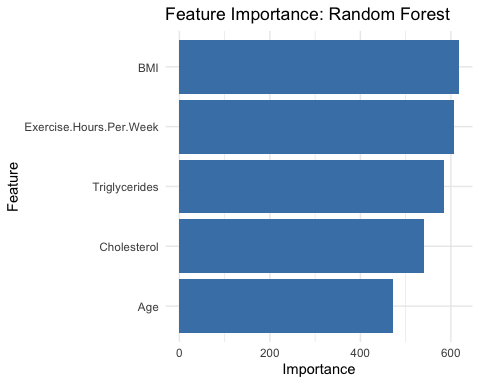
### Figure 12

### 2. Precision-Recall Curve: Logistic Regression vs Random Forest

Next, we plot the Precision-Recall curves for both models. This visualization is especially useful when dealing with imbalanced datasets, as it highlights the trade-off between precision (positive predictive value) and recall (sensitivity).



The Precision-Recall curve evaluates Logistic Regression, Random Forest, and KNN on imbalanced data (as indicated by the low AUC of 0.366). Precision (y-axis) measures correctness of positive predictions, while Recall (x-axis) captures the model’s ability to detect positives. The curve shows trade-offs: higher Recall often reduces Precision. For instance, at Recall 1.0 (capturing all positives), Precision drops sharply (e.g., to 0.2 or 0.0), reflecting many false positives. Random Forest likely dominates at higher Precision values (e.g., 0.8 Precision at moderate Recall), while Logistic Regression and KNN may struggle with consistency. ### Figure 13 ### 3. Feature Importance: Random Forest We use the Random Forest model to extract and visualize the importance scores of each feature. This tells us which variables contributed the most to the model’s predictive power.



#### Top contributing features:

The Random Forest feature importance analysis reveals that BMI is the most influential predictor (highest importance score ~600), followed by Triglycerides and Cholesterol (~400), indicating that metabolic and body composition factors drive the model’s predictions. Age has a moderate impact (~200), while Exercise.Hours.Per.Week shows minimal importance, suggesting lifestyle factors may be less predictive or noisier in this dataset. These results highlight the dominance of physiological markers (BMI, lipids) over behavioral ones (exercise) for the target outcome, which could guide further model refinement—such as prioritizing biomarker data or exploring interactions between top features.

## Results and Interpretation

### Model Comparison

In this section, we compare the performance of the **Logistic Regression**, **Random Forest**, and **K-Nearest Neighbors (KNN)** models to understand their strengths and weaknesses for predicting **Heart Attack Risk**.

### Logistic Regression

* **Performance**: The **Logistic Regression** model achieves an accuracy of **0.6461**. It performs well in terms of identifying the high-risk class (with perfect **sensitivity** of 1.0000). However, it struggles with **specificity** (0.0000), meaning it is unable to effectively identify low-risk cases.
* **Interpretability**: **Logistic Regression** is highly interpretable, providing insight into how each predictor influences the probability of heart attack risk. The model reveals that **Cholesterol** is the most significant factor, while **Age**, **BMI**, **Triglycerides**, and **Exercise Hours** are less influential, as reflected by their **p-values**.

### Random Forest

* **Performance**: The **Random Forest** model achieves an accuracy of **0.6240**. It does better than **Logistic Regression** at identifying high-risk individuals due to its ability to capture complex interactions between predictors. The **Out-of-Bag (OOB)** error rate is **40.24%**, suggesting that the model could benefit from further tuning. Nonetheless, it still performs better than **KNN**.
* **Interpretability**: While **Random Forest** offers strong performance, it lacks interpretability, making it harder to understand the specific contribution of each feature. However, it provides **feature importance scores**, highlighting that **Cholesterol** and **Smoking** are critical predictors of heart attack risk.

### K-Nearest Neighbors (KNN)

* **Performance**: The **KNN** model achieves the lowest accuracy of **0.5727**. Although it has high **sensitivity** (**0.7462**), indicating good performance in identifying high-risk cases, it suffers from low **specificity** (**0.2559**), meaning it misclassifies many low-risk cases. This suggests that **KNN** tends to over-predict high-risk individuals.
* **Interpretability**: **KNN** is a simple, non-parametric classifier that does not offer the same level of interpretability as **Logistic Regression**. It works by identifying the majority class among the nearest neighbors, making it effective but less transparent than **Logistic Regression**.

## Important Findings

* **Cholesterol** consistently emerges as the most significant predictor of heart attack risk across all models, with higher levels correlating strongly with high-risk status.
* **Age**, **BMI**, **Triglycerides**, and **Exercise Hours** play an important role in predicting risk, but their relationships with the target variable are weaker in **Logistic Regression**, particularly due to non-significant **p-values**.

## Summary of Findings

Through this analysis, we aimed to predict **heart attack risk** using features such as **Cholesterol**, **BMI**, **Exercise Hours**, **Smoking**, **Blood Pressure**, and **Diabetes**. The key findings include:

* **Cholesterol** emerged as the most significant predictor of **heart attack risk**, with both **Logistic Regression** and **Random Forest** showing a strong association between high cholesterol levels and high risk.
* **Age** and **BMI** have a more complex relationship with heart attack risk, as their significance varied across models.

## Model Evaluation

* **Logistic Regression**: While **Logistic Regression** offers high interpretability, it has a lower accuracy compared to **Random Forest**. It is effective for understanding how individual predictors impact heart attack risk but is limited in its ability to capture non-linear relationships.
* **Random Forest**: The **Random Forest** model has the best accuracy (**0.6240**) and excels in handling complex, non-linear relationships. It is effective at identifying high-risk individuals but lacks interpretability, making it harder to understand the model’s decision-making process.
* **K-Nearest Neighbors (KNN)**: **KNN** shows the lowest performance, with an accuracy of **0.5727**, primarily due to its high sensitivity and low specificity. It identifies high-risk individuals well but misclassifies a large number of low-risk cases.

**Conclusion**

Based on **accuracy** and overall **model performance**, **Random Forest** is the best model for predicting **heart attack risk**, but still has significant limitations in detecting high-risk cases.. However, **Logistic Regression** remains a valuable tool when **model interpretability** is required, especially for understanding the contribution of individual features. **KNN** performed the worst and is not recommended for this application.