HEART ATTACK RISK PREDICTION AND ANALYSIS IN INDIAN POPULATION

Project Report submitted in partial fulfilment of the requirements
for the award of Degree of Master of Science in Statistics
with Specialization in Data Analytics

COURSE NAME : M.Sc. Statistics with Specialization in Data Analytics

UNIVERSITY COLLEGE, THIRUVANANTHAPURAM
UNIVERSITY OF KERALA
2023-2025

CONTENTS

CONTENTS **Page No.**

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cardiovascular diseases, particularly heart attacks, are one of the leading causes of mortality worldwide, and the burden is rapidly increasing in developing countries like India. With changing lifestyles, urbanization, and increased exposure to risk factors such as smoking, poor diet, stress, diabetes, and hypertension, the Indian population has become increasingly vulnerable to heart-related conditions. Despite this, early detection and prevention strategies often remain underutilized due to a lack of personalized, data-driven risk assessments.

This project focuses on predicting heart attack risk and identifying key contributing factors using data from the Indian population. By applying statistical techniques and machine learning models, the aim is to uncover meaningful patterns and insights that can help in early intervention and better risk stratification. The study explores clinical parameters like blood pressure, cholesterol levels, diabetes status, and lifestyle attributes such as smoking, alcohol consumption, and physical activity. It also investigates regional disparities by incorporating demographic variables such as state of residence and access to healthcare.

The motivation behind this project lies in developing a data-informed approach that not only aids in individual risk prediction but also informs public health policies. By combining Exploratory Data Analysis (EDA), hypothesis testing, and classification models, this study strives to provide a holistic view of heart attack risk factors prevalent in India. The outcomes can support clinicians in understanding high-risk profiles and empower individuals to make informed health decisions(refer to Prabhakaran et al.,2016).

1.2 Objectives of the Study

1. **To identify and analyze key clinical and lifestyle factors** (such as blood pressure, cholesterol levels, diabetes, obesity, stress, and smoking) associated with heart attack risk in the Indian population.
2. **To perform statistical hypothesis testing** (e.g., chi-square tests, logistic regression, ANOVA) to determine significant predictors of heart attack occurrence.
3. **To build and evaluate predictive models** (e.g., logistic regression, decision trees) for estimating the likelihood of heart attacks using structured health data.
4. **To explore geographical and demographic variations** in heart attack risk across different Indian states and age groups.

5. **To segment the population into risk clusters** using unsupervised learning techniques like K-Means clustering for enhanced group profiling.

1.3 Dataset

Kaggle Dataset: Heart Attack Risk & Prediction Dataset In India

1.4 About Dataset

Cardiovascular diseases (CVDs) are the leading cause of death in India, with heart attacks (myocardial infarctions) accounting for a significant portion. India has a higher heart disease burden than many other nations, with cases occurring at younger ages compared to Western countries. This dataset incorporates key medical and lifestyle risk factors such as diabetes, hypertension, obesity, smoking, air pollution exposure, and healthcare access.

1.5 Summary

Using a dataset representative of the Indian population, this project carried out end-to-end analysis and prediction of heart attack risk. Through data preprocessing, feature selection, model training, and performance evaluation, it was found that factors such as **age, diabetes, LDL levels, blood pressure, and stress** have a significant impact on risk levels.

Machine learning models were compared, and Logistic Regression with balanced class weights showed interpretable results, while Random Forests were used for deeper insight. Clustering analysis revealed natural risk groups among patients, emphasizing the role of unsupervised learning in medical profiling.

Overall, this study provides a robust framework for heart attack risk prediction, highlighting the importance of preventive care in the Indian context.

CHAPTER 2

REVIEW OF LITERATURE

Cardiovascular diseases (CVDs), particularly heart attacks, have become a major global health concern and a leading cause of death in India. Research over the last few decades has focused on identifying significant risk factors and developing predictive models to enable early diagnosis and intervention. The integration of statistical methods and machine learning algorithms has significantly advanced the precision and interpretability of heart disease prediction systems.

The early foundation for heart disease risk analysis was established using classical statistical techniques. Gupta and Kapoor (2007) emphasized the importance of descriptive and inferential statistics in understanding the influence of demographic and physiological variables such as age, blood pressure, and cholesterol levels. Gupta (1990) further contributed to the statistical framework through regression analysis and hypothesis testing, enabling researchers to evaluate the impact of individual risk factors.

In the context of clinical risk prediction, logistic regression has long been a preferred method due to its interpretability and ability to model binary outcomes such as the presence or absence of heart disease. Studies utilizing logistic regression (e.g., Detrano et al., 1989) identified cholesterol, blood pressure, and age as prominent contributors to heart attack risk. However, with the growing availability of health data, more complex models were required to capture non-linear relationships and interactions between variables.

Recent advancements in machine learning have revolutionized predictive modeling in healthcare. Algorithms such as Random Forest, Support Vector Machines (SVM), AdaBoost, and Gradient Boosting have been increasingly applied in heart disease prediction. Gudadhe et al. (2010) demonstrated the potential of decision tree-based models to outperform traditional techniques, especially in handling large datasets with heterogeneous features. Similarly, Dey et al. (2018) compared multiple machine learning models and found Random Forest to be the most effective in predicting heart disease outcomes due to its ability to manage feature importance and reduce overfitting.

In addition to predictive accuracy, the interpretability of machine learning models is crucial in healthcare applications. Several studies have utilized feature importance techniques to identify key variables contributing to heart attack risk. Variables such as LDL and HDL cholesterol levels, systolic and diastolic blood pressure, age, and stress level consistently ranked high across different models. These findings reinforce existing clinical knowledge and provide insights for targeted public health interventions.

Moreover, clustering techniques like K-means have been used to segment patients into distinct groups based on shared characteristics. Research has shown that unsupervised learning can help uncover hidden patterns in data, revealing population subgroups with elevated heart attack risk. The Elbow method has proven effective in determining optimal cluster numbers, while cluster-wise risk analysis enables public health planners to focus on high-risk demographics.

Several Indian studies have highlighted the unique challenges and risk profiles in the subcontinent. According to Prabhakaran et al. (2018), the Indian population exhibits a higher propensity for heart disease at younger ages, often linked to urban lifestyle factors, stress, and lack of physical activity. These studies emphasize the need for population-specific models rather than relying solely on Western datasets and benchmarks.

Another notable challenge in Indian studies is the availability and quality of healthcare access data. Rural populations often face limited access to timely care and preventive screenings, which amplifies the importance of early risk prediction using accessible variables such as age, blood pressure, and lifestyle indicators. Studies by Yusuf et al. (2004) suggest that incorporating environmental and socio-economic factors—like air pollution and healthcare access—improves the relevance of prediction models in developing countries.

The integration of electronic health records (EHRs), wearable sensor data, and patient-reported metrics has expanded the horizon of data-driven heart disease prediction. Modern systems use ensemble models and feature engineering to harness this multidimensional data for real-time risk scoring. Research by Sharma et al. (2021) implemented ensemble learning methods combining Random Forest and AdaBoost, reporting high sensitivity and accuracy in classifying high-risk individuals.

In summary, the literature reveals a progressive shift from traditional statistical methods to advanced machine learning approaches in heart attack risk prediction. The fusion of statistical rigor with algorithmic modeling enhances both the interpretability and predictive strength of such systems. Although global advancements are notable, there remains a need for localized research tailored to the Indian population's unique risk factors and healthcare landscape. This study builds on these gaps by integrating both statistical analysis and machine learning techniques to model heart attack risk, with a focus on interpretability, accuracy, and practical implications for health interventions in India

CHAPTER 3

METHODOLOGY

This study followed a structured approach combining Exploratory Data Analysis (EDA), statistical testing, and predictive modeling to identify key risk factors and predict heart attack risk in the Indian population.

3.1 Data Collection and Preprocessing

- The dataset was sourced from a publicly available health dataset, containing 10,000 records and 26 clinical, lifestyle, and demographic variables.
- Missing values were handled using appropriate imputation techniques or removed if minimal.
- Categorical variables (e.g. State, Gender) were encoded using one-hot encoding or label encoding.
- Features were scaled using StandardScaler for model compatibility(refer to Deo R. C, 2015).

3.2 Exploratory Data Analysis (EDA)

- Univariate and bivariate analysis were conducted to understand distributions and relationships.
- Visualizations included bar plots, boxplots, histograms, and heatmaps.
- Correlation analysis was used to identify initial relationships among variables.

3.3 Statistical Analysis

- Hypothesis testing

3.3.1 Chi-square tests for categorical variables vs. heart attack risk(target variable).

Purpose:
The chi-square test was used to examine whether there is a statistically significant association between two categorical variables—specifically, between categorical risk factors

(e.g., smoking, gender, alcohol consumption) and the binary outcome variable: heart attack risk (low or high).

Hypotheses:

- **Null hypothesis (H_0)**: The categorical variable is independent of heart attack risk.
- **Alternative hypothesis (H_1)**: The categorical variable is associated with heart attack risk.

Test Statistic:

$$\chi^2 = \frac{\sum (O - E)^2}{E}$$

Where:

- O is the observed frequency
- E is the expected frequency under the assumption of independence

Interpretation:
A **p-value** less than the chosen significance level (typically 0.05) indicates that the variable is significantly associated with heart attack risk.

Application Example:
Applied to variables like **Smoking**, **Gender**, and **Alcohol Consumption** to evaluate their association with heart attack risk.

3.3.2. Independent Samples t-Test

Purpose:
The independent samples t-test was used to compare the means of a continuous variable (e.g., LDL level, age) between two groups—typically those with high and low heart attack risk.

Hypotheses:

- Null hypothesis (H_0): The mean value of the variable is the same for both risk groups.
- Alternative hypothesis (H_1): There is a significant difference in the mean values between the groups.

Test Statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{X}_1, \bar{X}_2$ = sample means
- s_1^2, s_2^2 = sample variances
- n_1, n_2= sample sizes

Assumptions:

- The two groups are independent.
- The variable is approximately normally distributed in each group.
- Variances are equal (can be tested using Levene's test).

Interpretation:
A low p-value suggests a significant difference in the means, indicating that the variable may play a role in heart attack risk.

3.3.3. One-Way ANOVA (Analysis of Variance)

Purpose:
ANOVA was used to determine whether the mean of a continuous variable (like age) differs across more than two groups. In this project, it was applied to compare mean heart attack risk scores across **age groups**.

Hypotheses:

- **Null hypothesis (H_0)**: The mean value is equal across all groups.
- **Alternative hypothesis (H_1)**: At least one group mean is significantly different.

Test Statistic:

$$F = \frac{Between\ group\ variability}{Within\ group\ variability} = \frac{MS_{between}}{MS_{within}}$$

Assumptions:

- Each group sample is independent.
- Each group has a normal distribution.
- Homogeneity of variances (equal variances across groups).

Interpretation:
If the **F-statistic** yields a p-value less than 0.05, it implies that not all group means are equal.

- Results were used to assess the statistical significance of features(refer to Gupta S. C.,2007).

3.4 Feature Engineering

- New features like Age_Group and interaction terms (e.g., LDL × Diabetes) were created based on domain knowledge.
- Redundant or non-informative features were removed.

3.5 Modeling and Evaluation

- **Machine Learning Models**:
    - Logistic Regression (with and without class balancing)
    - Decision Trees and Random Forests .
- Models were trained on an 80-20 train-test split.
- Evaluation metrics included accuracy, precision, recall, F1-score, and confusion matrix(refer to Lundberg S. M., & Lee S. I.,2017).

3.5.1 Logistic Regression

Purpose:
Logistic regression is used for binary classification problems where the outcome variable is categorical (e.g., heart attack: *yes* or *no*).

Statistical Basis:

Logistic regression models the probability that a given input belongs to a particular category using the logistic (sigmoid) function:

$$P(Y=1)=\frac{1}{1+\exp\left(-\left(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n\right)\right)}$$

Where:

- Y is the binary response variable (Heart_Attack_Risk)
- $X_1, X_2, \ldots, X_n$ are predictor variables (like cholesterol, age)
- $\beta_0, \beta_1, \ldots, \beta_n$ are model coefficients(refer to Gupta S. C et al.,2007)

Use in this Study:

Logistic regression was employed to identify significant predictors of heart attack risk and to estimate how each variable (e.g., cholesterol levels, age, stress) affects the probability of experiencing a heart attack.

3.5.2 Decision Tree Classifier

Purpose:

A Decision Tree is a flowchart-like model that splits the dataset into subsets based on feature values, ultimately predicting a target variable.

Principle:

The tree selects features that best split the data using criteria like **Gini Impurity** or **Entropy (Information Gain)**.

For Gini Index:

$$\text{Gini}=1-\sum p_i^2$$

Where p_i is the probability of class i.

Use in this Study:

A Decision Tree model was trained on clinical and demographic data to create a simple, interpretable model for identifying high-risk patients(refer to Shouman M,2012).

3.5.3 Random Forest Classifier

Purpose:

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve predictive accuracy and control overfitting(refer to Breiman L., 2001)..

Statistical Idea:

Each tree in the forest is trained on a bootstrap sample of the data, and at each split, only a random subset of features is considered.

Final prediction is made by:

- **Classification:** Majority voting among trees.
- **Regression:** Averaging predictions.

Use in this Study:

Random Forest was used to improve accuracy over single models and to determine **feature importance** — i.e., which factors contribute most to heart attack risk.

3.6 Clustering (Unsupervised Learning)

- K-Means clustering was used to identify natural patient groupings based on risk factors.
- Elbow method guided the selection of optimal k (refer to Jain A. K.,2010).

3.7 Interpretation

- Coefficients from logistic regression were interpreted to assess feature importance.

CHAPTER 4

ANALYSIS

4.1 Introduction

This chapter presents the detailed exploratory data analysis (EDA) and modeling results aimed at understanding and predicting heart attack risk within the Indian population. The analysis begins by examining the distribution of clinical, lifestyle, and demographic variables, identifying patterns and potential risk factors associated with heart disease.

Both statistical techniques and machine learning methods were applied to uncover relationships between variables and the likelihood of heart attack occurrence. EDA helped to detect data imbalances, outliers, and significant trends, while statistical tests such as Chi-square, t-tests, and ANOVA validated associations between predictors and heart attack risk.

Subsequently, predictive models including Logistic Regression, Decision Trees, and Random Forests were developed. Each model's performance was evaluated using classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to determine their effectiveness in classifying heart attack risk.

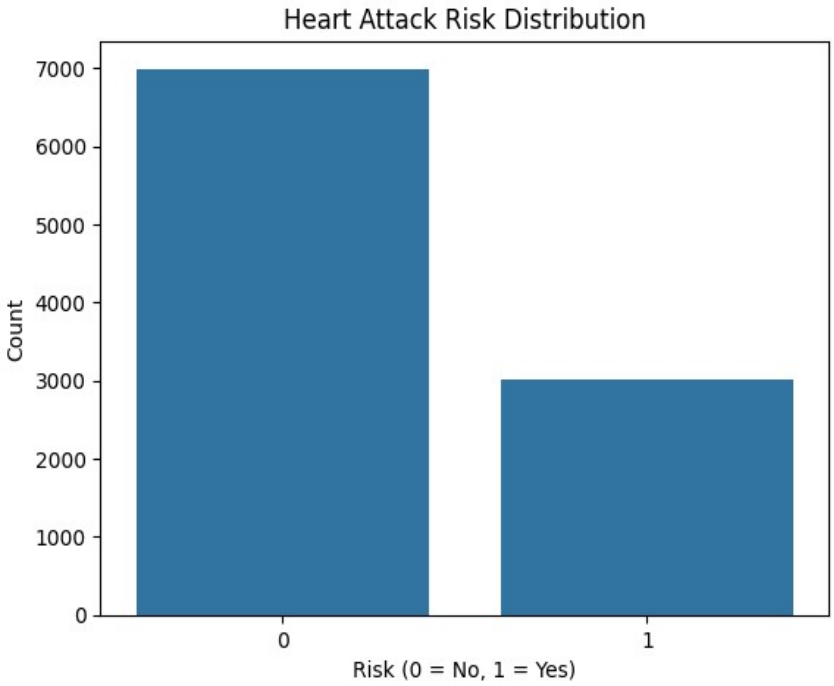
Additionally, clustering techniques were used to explore natural groupings among patients, revealing risk profiles that could be useful in clinical segmentation. The results from all analyses are interpreted in the context of healthcare relevance, especially considering the unique characteristics of the Indian demographic.

4.2 Distribution of Heart Attack Risk

To gain an initial understanding of how heart attack risk is distributed in the dataset, visualizations such as bar plots were used. These diagrams provided a clear overview of the proportion of individuals classified as high-risk versus low-risk. Such representations help in identifying class imbalance and guiding appropriate preprocessing or modeling strategies.

According to **Diagram 4.1**, it shows that approximately 69.93% of individuals are categorized as not at risk (0), while 30.07% are at risk (1). This class imbalance is important to consider during model training to ensure balanced performance.

Diagram 4.1

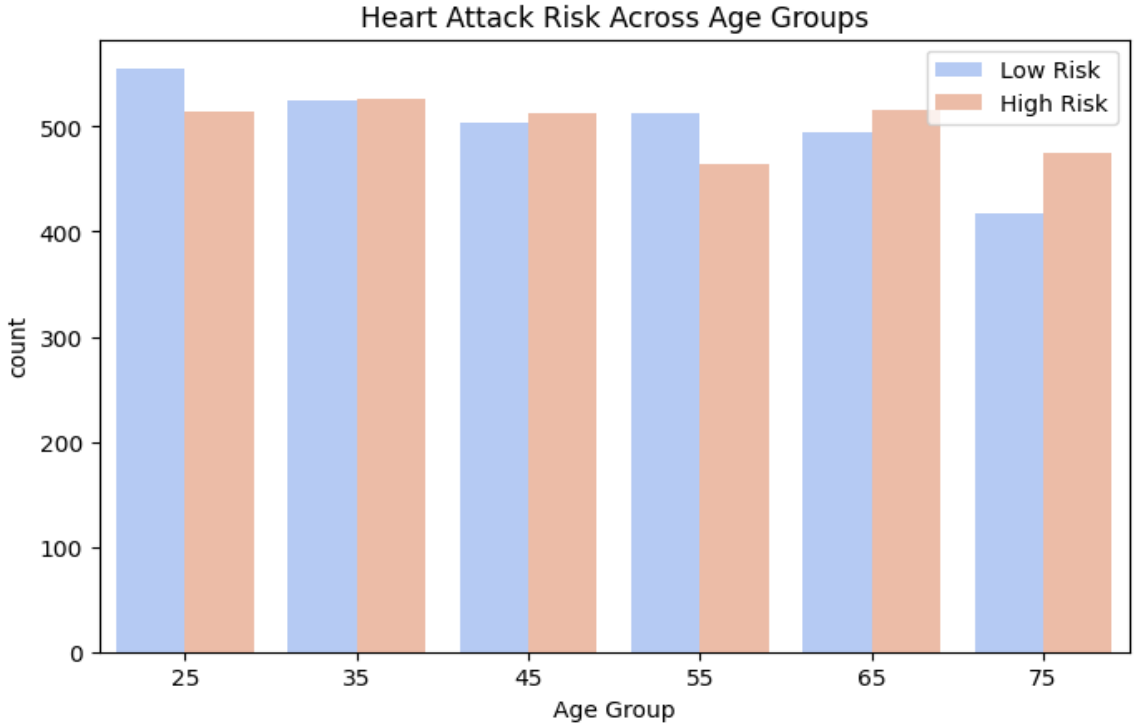



```
Heart_Attack_Risk
0    69.93
1    30.07
```

4.3 Age Distribution

Bar Chart to Visualize Age Distribution

Diagram 4.2

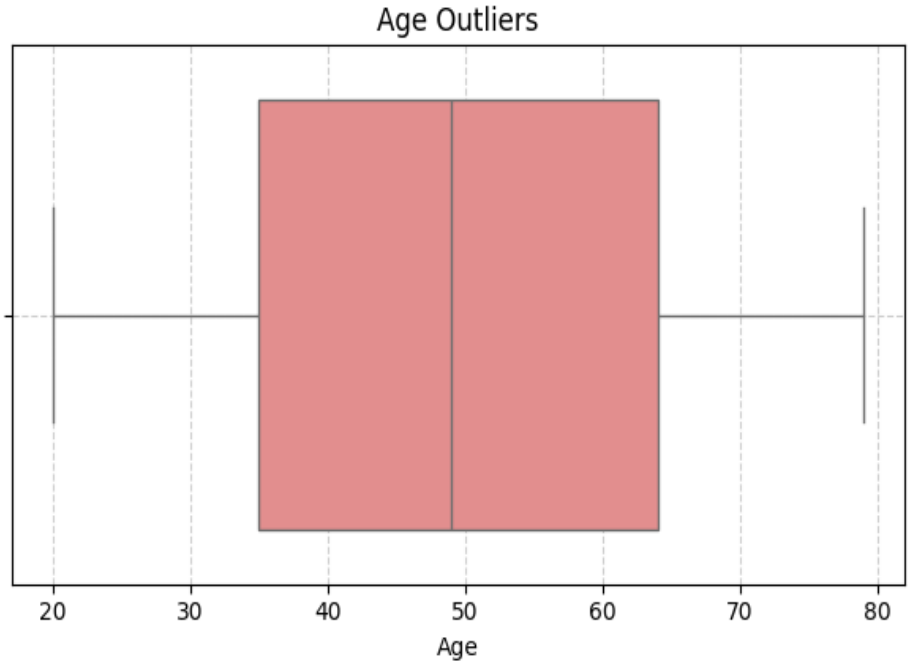


Heart attack risk analysis was conducted across a broad age spectrum, enabling meaningful comparisons across different life stages.

The bar chart (**Diagram 4.2**) illustrates heart attack risk across different age groups. It shows a gradual increase in high-risk individuals in older age brackets, especially beyond 65 years, indicating age as **a significant factor influencing heart attack** susceptibility.

Box plot to identify age outliers

Diagram 4.3



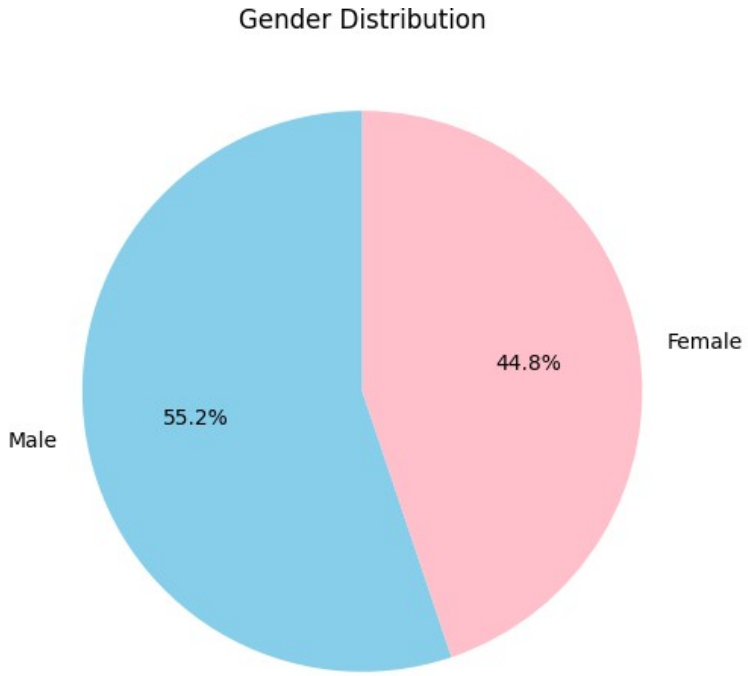
The box plot displays the distribution of age and highlights the absence of significant outliers. Most individuals fall between ages 30 and 70, with a median age around 50, indicating a fairly balanced age spread among the participants.

4.4 Gender Distribution

Pie Chart for Gender Distribution

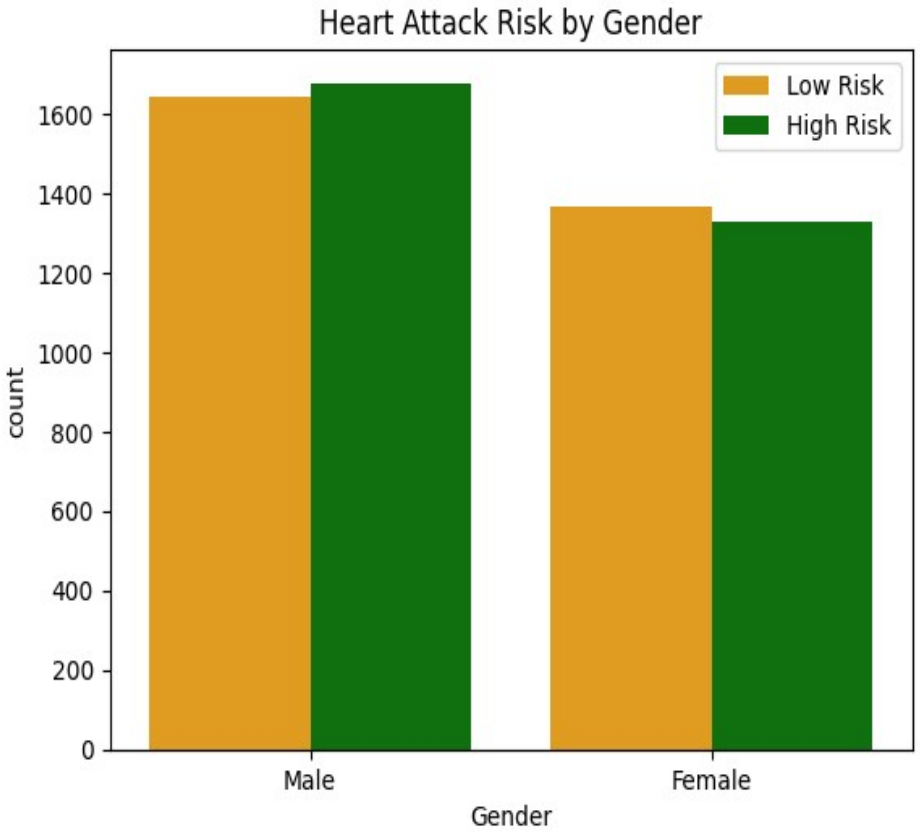
The pie chart illustrates the gender distribution within the dataset. Males constitute a slightly higher proportion at 55.2%, while females represent 44.8%, indicating a moderately balanced gender representation.

Diagram 4.4



Gender-Wise Heart Attack Risk

Diagram 4.5

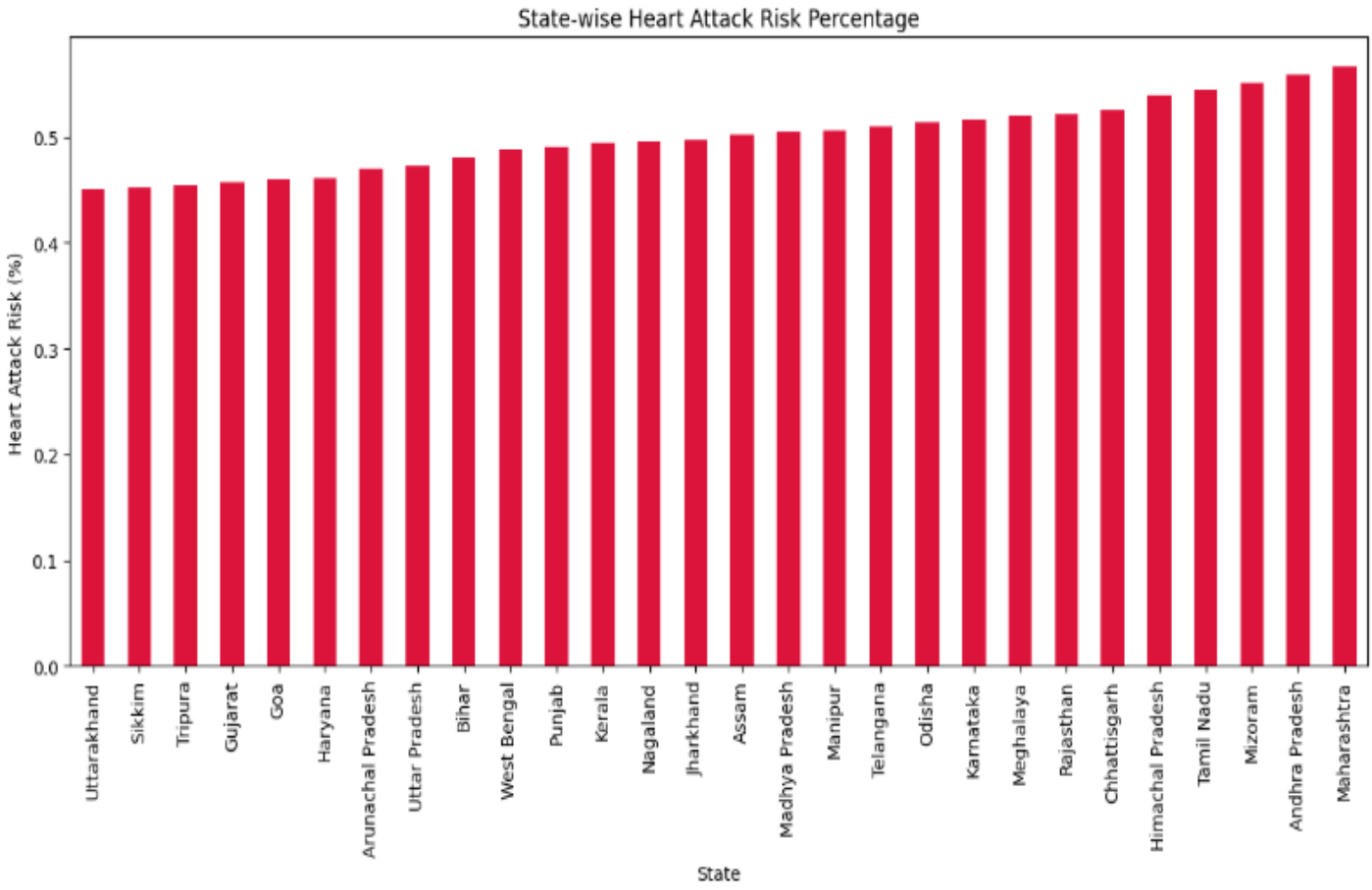


The bar chart illustrates the distribution of heart attack risk among males and females. It is evident that males have a higher count in both low-risk and high-risk categories compared to females. Among males, the number of high-risk individuals slightly exceeds the number of low-risk individuals, indicating a relatively greater vulnerability to heart attacks. In contrast, females have slightly more individuals in the low-risk category than in the high-risk group. This suggests that, overall, females may have a marginally lower risk profile for heart attacks. The visual also highlights that **gender plays a significant role in risk distribution**, with **men exhibiting a consistently higher incidence of heart attack risk**. These insights are important for tailoring gender-specific preventive strategies.

4.5 State-Wise Heart Attack Risk

The bar chart displays the heart attack risk percentages across different Indian states. Maharashtra shows the highest risk percentage, while Uttarakhand has the lowest. The chart highlights regional variations in cardiovascular health risks.

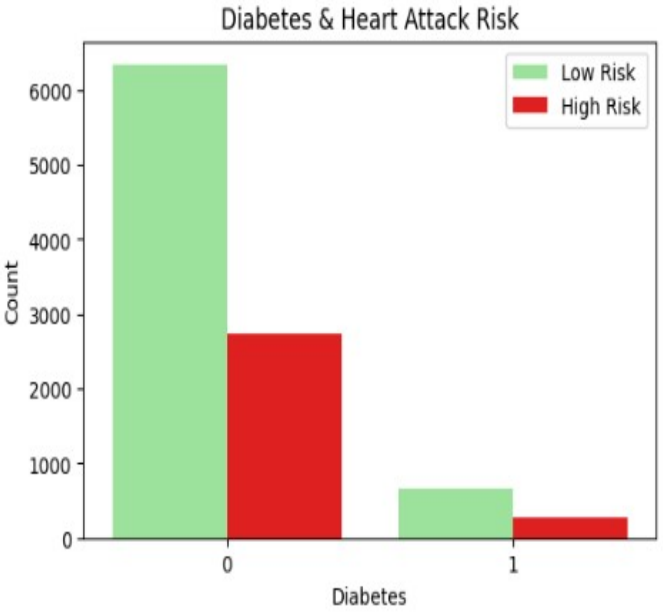
Diagram 4.6

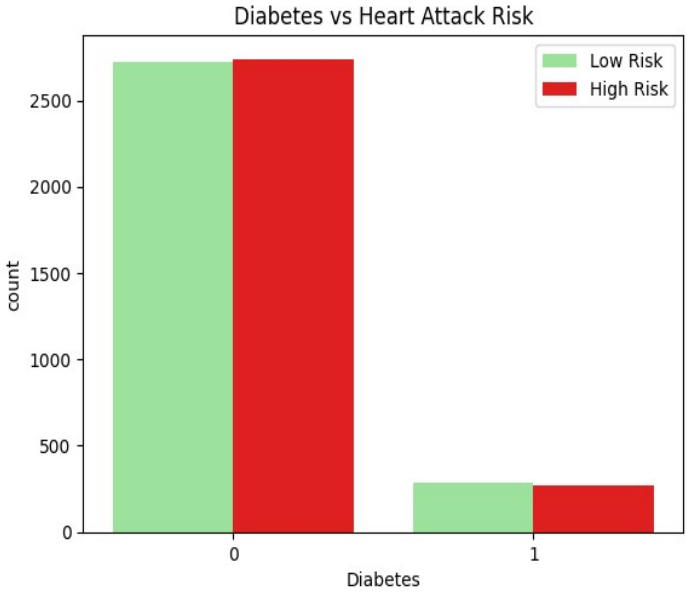


4.6 Impact of Lifestyle Factors on Heart Attack Risk

4.6.1 Diabetes vs Heart Attack Rise

Diagram 4.7





Before Balancing the Data

After Balancing the Data

To analyze the influence of diabetes on heart attack risk, two visualizations were generated — one before and one after balancing the dataset.

The first graph (unbalanced) reflects the real-world distribution, where non-diabetic individuals significantly outnumber diabetic individuals. This imbalance could potentially bias any predictive model toward the majority class, under-representation the risk faced by diabetic individuals.

To address this, the dataset was balanced using appropriate techniques to ensure equal representation of diabetic and non-diabetic individuals. The second graph (balanced) provides a clearer and more unbiased view of how diabetes contributes to heart attack risk.

By comparing both plots, it becomes evident that **balancing the data enhances the analytical power and allows for more accurate interpretation of diabetes as a risk factor in predictive modeling.**

Inference:

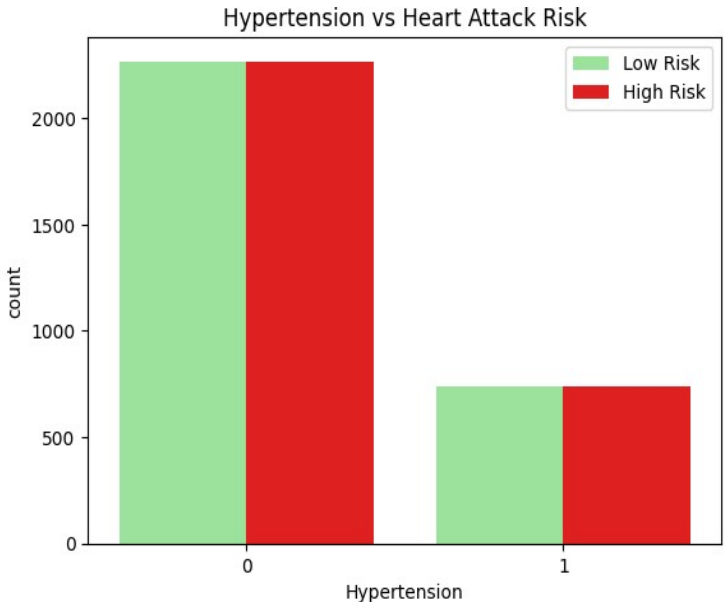
This chart illustrates the relationship between diabetes and heart attack risk. Individuals without diabetes (0) show a significantly higher count, with both low and high heart attack risks being almost equal. Among diabetic individuals (1), although the total count is lower, the high-risk category is nearly as prevalent as the low-risk one. This suggests that diabetes may **contribute to an elevated risk of heart attacks.**

4.6.2 Hypertension vs Heart Attack Risk

The bar chart illustrates the association between hypertension status and heart attack risk. Individuals without hypertension (coded as 0) show a high frequency for both low and high heart attack risk categories, indicating a balanced distribution. However, among those with hypertension (coded as 1), although the total count is lower, the number of high-risk individuals remains significant and nearly matches the number of low-risk individuals.

This pattern highlights that people with hypertension are proportionally more prone to high heart attack risk compared to those without hypertension. **It emphasizes hypertension as a key lifestyle-related factor that could contribute significantly to cardiovascular risk and underscores the need for early detection and management of blood pressure.**

Diagram 4.8

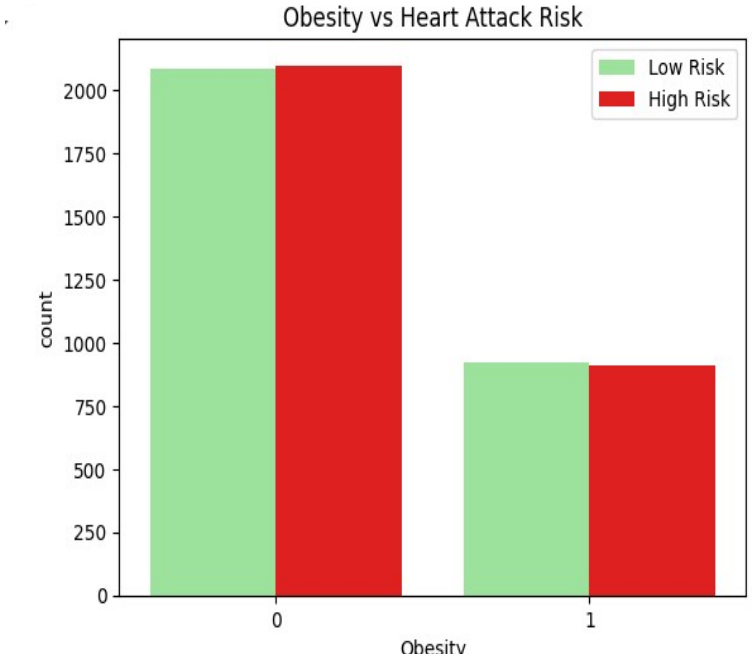


4.6.3 Obesity vs Heart Attack Risk

The bar chart illustrates the relationship between obesity and heart attack risk. Among non-obese individuals (label 0), both low and high risk counts are relatively high and nearly equal. In contrast, among obese individuals (label 1), the count is lower overall but shows a balanced distribution between risk levels.

This suggests that while obesity may be associated with heart attack risk, the distribution **does not show a strong imbalance, indicating the need to consider additional factors for clearer insights.**

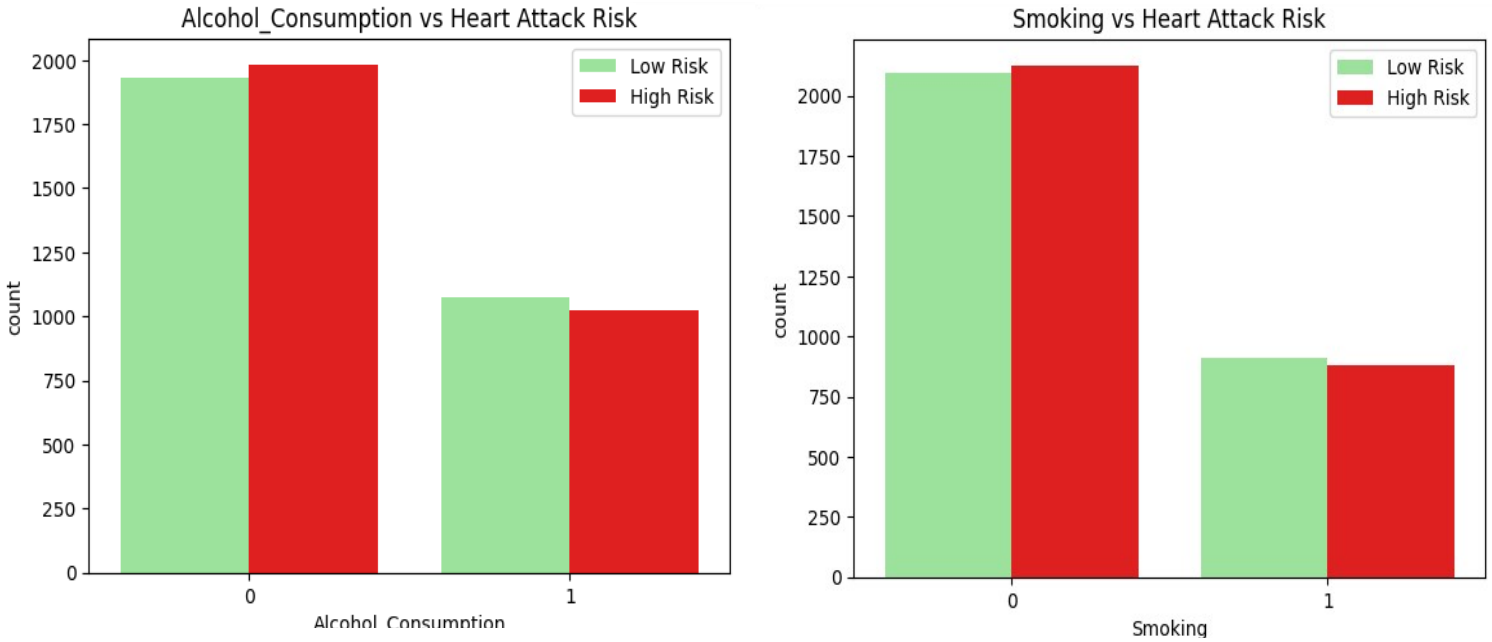
Diagram 4.9



4.6.4 Smoking, Alcohol Consumption vs Heart Attack risk

The below graphs illustrates the relationship between smoking status, Alcohol consumption vs heart attack risk. This suggests **smoking may be a contributing factor but not the sole determinant of heart attack risk** and **alcohol consumption alone may not strongly differentiate heart attack risk in this dataset.**

Diagram 4.10

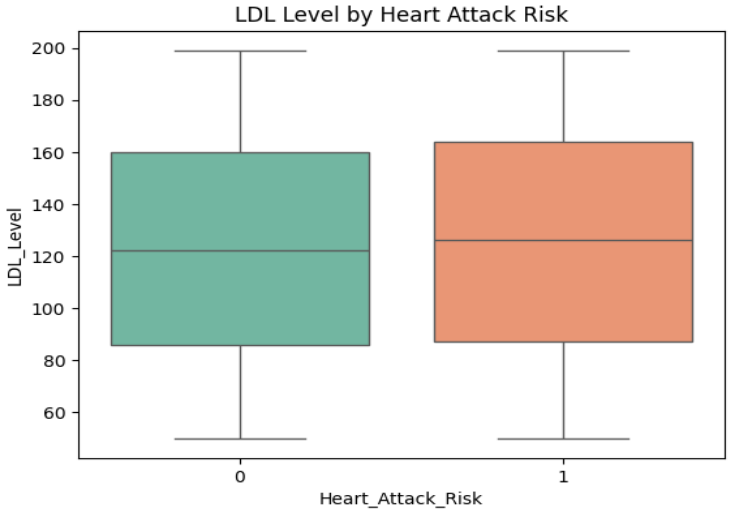


Overall, the analysis highlights **smoking, diabetes, and hypertension** as the most impactful lifestyle-related contributors to heart attack risk. While alcohol consumption and obesity also show some correlation, their influence appears to be comparatively weaker in this particular dataset. Adopting healthier lifestyle choices in these areas may help reduce the likelihood of heart-related issues.

4.7 Clinical Features Influencing Heart Attack risk

4.7.1 LDL Level vs Heart Attack Risk

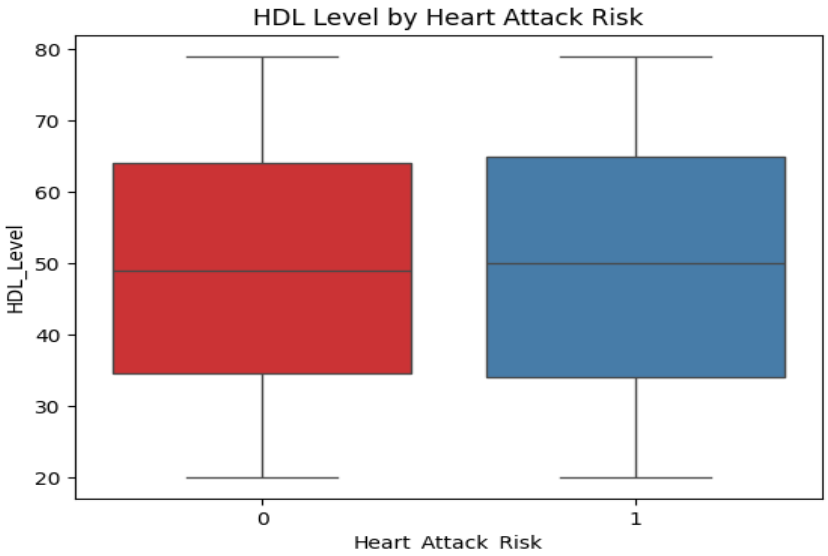
Diagram 4.11



The boxplot above compares LDL cholesterol levels between individuals with low (0) and high (1) heart attack risk. Both groups show similar distributions with overlapping interquartile ranges, suggesting that LDL levels are slightly higher in the high-risk group but not significantly different. The median LDL level appears marginally elevated for those at higher risk, reinforcing that **elevated LDL may contribute to heart attack risk**, though it may not be a strong independent predictor on its own in this dataset.

4.7.2 HDL Level vs Heart Attack Risk

Diagram 4.12



This box plot depicts the distribution of HDL (High-Density Lipoprotein) levels for individuals categorized by Heart Attack Risk. The median HDL levels for both low-risk and high-risk groups appear quite similar, with overlapping interquartile ranges and nearly identical overall spreads. This suggests that **HDL levels alone may not significantly distinguish** between low and high heart attack risk in this dataset. However, consistently lower **HDL levels could still contribute to overall cardiovascular risk when combined with other factors.**

>> **Statistical Significance for a more detailed conclusion**

Logistic Regression Results Summary:

- **Model**: Binary logistic regression
- **Dependent variable**: Heart_Attack_Risk
- **Predictors**: LDL_Level, HDL_Level

Key Findings:

1. LDL_Level

- **Coefficient**: 0.0013
- **p-value**: 0.023 (Statistically significant at the 5% level)
- **Interpretation**: As **LDL (bad cholesterol)** increases, the **odds of heart attack risk increase slightly**. This aligns with clinical expectations.

2. HDL_Level

- **Coefficient**: 0.0010
- **p-value**: 0.495 (Not statistically significant)
- **Interpretation**: **HDL (good cholesterol)** has **no significant effect** on heart attack risk in this model.

The logistic regression model suggests that LDL cholesterol is a **significant positive predictor** of heart attack risk, while HDL cholesterol does **not show a statistically significant association**. However, the model's overall predictive power is low, indicating that other variables may be necessary for a better risk prediction.

Diagram 4.13

```
Optimization terminated successfully.
         Current function value: 0.692675
         Iterations 3
                           Logit Regression Results
==============================================================================
Dep. Variable:      Heart_Attack_Risk   No. Observations:                 6014
Model:                          Logit   Df Residuals:                     6011
Method:                           MLE   Df Model:                            2
Date:                Tue, 06 May 2025   Pseudo R-squ.:               0.0006810
Time:                        18:08:44   Log-Likelihood:                -4165.7
converged:                       True   LL-Null:                       -4168.6
Covariance Type:            nonrobust   LLR p-value:                   0.05850
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2172      0.106     -2.049      0.040      -0.425      -0.009
LDL_Level      0.0013      0.001      2.266      0.023       0.000       0.003
HDL_Level      0.0010      0.001      0.682      0.495      -0.002       0.004
==============================================================================
```

4.7.3 Heart Attack Risk by Blood Pressure Category

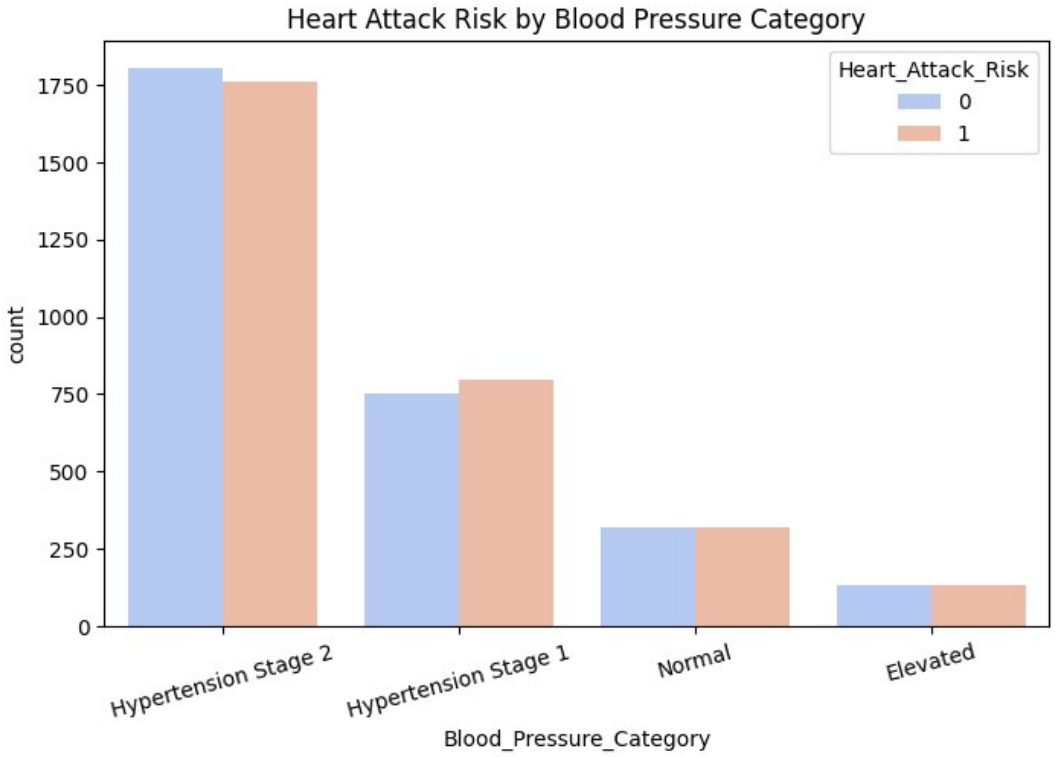
This bar chart (**Diagram 4.13**) displays the distribution of heart attack risk across different blood pressure categories:

- Individuals in **Hypertension Stage 2** have the highest count in both low and high risk categories, with nearly equal distribution. **This suggests a strong association between severe hypertension and heart attack risk.**
- **Hypertension Stage 1** also shows a high number of cases, slightly more for the high-risk group, indicating **moderate hypertension is also a contributor.**
- For those with Normal or Elevated blood pressure, the heart attack risk is significantly lower, and counts are much smaller.

Conclusion:

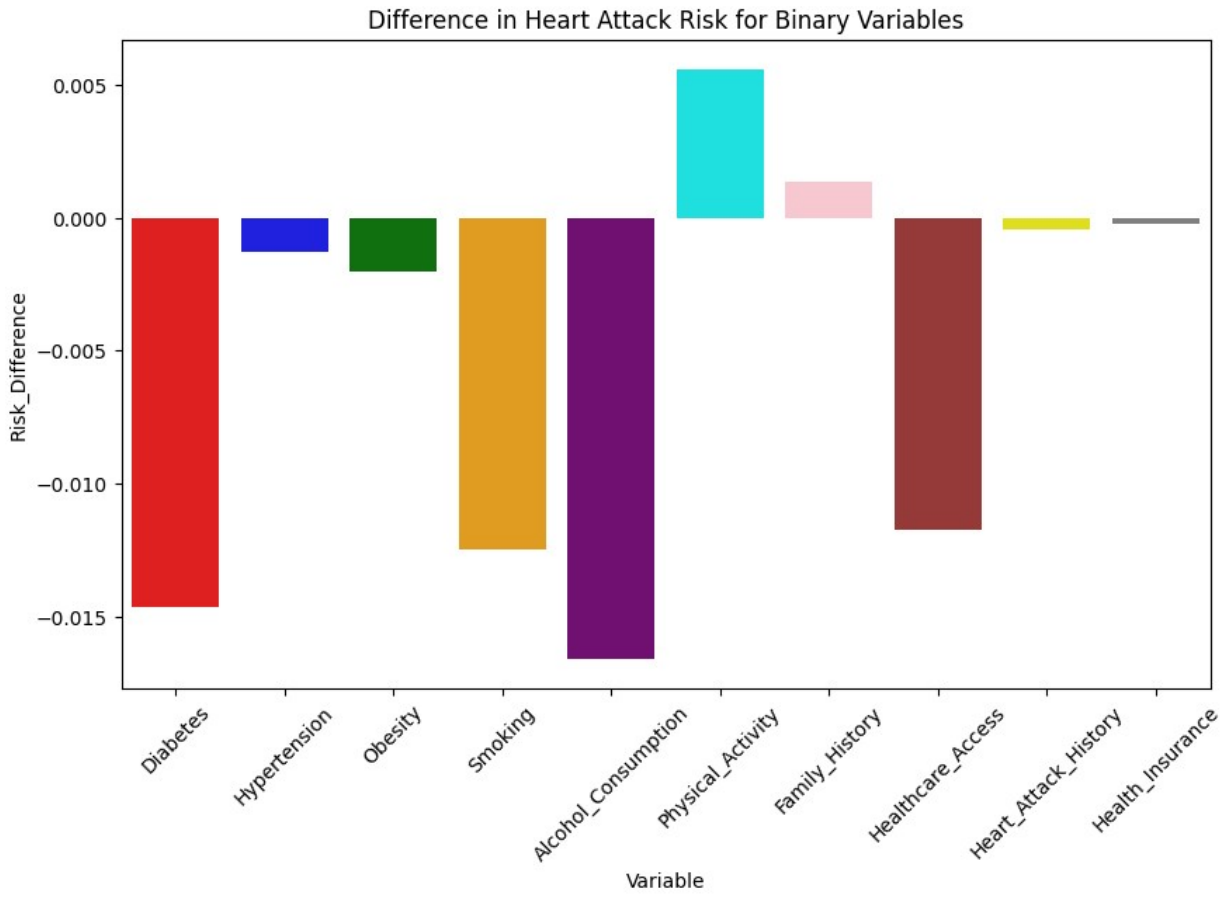
Higher stages of hypertension (especially Stage 2) are clearly associated with a greater likelihood of heart attack, highlighting **blood pressure as a critical factor in cardiovascular risk assessment.**

Diagram 4.14



4.8 Difference in Heart Attack Risk for Binary Variables

Diagram 4.15



This bar chart illustrates how various binary clinical and lifestyle variables influence heart attack risk by measuring the risk difference between individuals with and without each condition or behaviour.

- Negative bars (below 0) indicate variables associated with a higher heart attack risk.
    - **Alcohol Consumption and Diabetes** show the largest negative differences, suggesting a stronger association with increased heart attack risk.
    - Other negatively associated factors include **Smoking, Healthcare Access, Obesity, and Hypertension.**
- Positive bars (above 0) imply a lower heart attack risk for those with the characteristic.

- Physical Activity and Family History (unexpectedly) show small positive differences.
- These may reflect confounding factors or sample-specific patterns.

- Minimal impact is seen from Heart Attack History and Health Insurance, suggesting little to no direct correlation in this dataset.

Thus, Risk difference analysis highlights key binary factors contributing to heart attack risk. Elements such as alcohol consumption and smoking, along with diabetes and hypertension, were linked to higher risk, while physical activity appeared modestly protective.

4.9 Statistical Significance Testing

4.9.1 Chi-square Test (Categorical vs Categorical)

Chi-Square test provide insight into whether specific categorical variables are statistically significant in their association with heart attack risk.

```
Gender: p-value = 0.4341 (Not significant)
Diabetes: p-value = 0.3733 (Not significant)
Hypertension: p-value = 0.9223 (Not significant)
Obesity: p-value = 0.8598 (Not significant)
Smoking: p-value = 0.2200 (Not significant)
Alcohol_Consumption: p-value = 0.0881 (Not significant)
Physical_Activity: p-value = 0.5656 (Not significant)
Air_Pollution_Exposure: p-value = 0.5301 (Not significant)
Family_History: p-value = 0.9092 (Not significant)
Healthcare_Access: p-value = 0.2450 (Not significant)
Heart_Attack_History: p-value = 0.9967 (Not significant)
Health_Insurance: p-value = 0.9995 (Not significant)
```

Key Observations:

- None of the tested variables (e.g., **Gender, Diabetes, Hypertension, Smoking, Alcohol Consumption**, etc.) show a **p-value below 0.05**, indicating that:

  There is no statistically significant association between these categorical features and heart attack risk at the 5% level.

- For example:
  - **Diabetes**: p = 0.3733 ⟶ Not significant
  - **Smoking**: p = 0.2200 ⟶ Not significant

- **Alcohol Consumption**: p = 0.0881 ⟶ Approaching significance, but still not below 0.05

Interpretation:

While many of these variables are clinically known to influence heart health, the **Chi-Square test** suggests that in this dataset, they do **not show a statistically significant relationship** with heart attack risk. This could be due to sample size or overlapping effects with other variables.

Visual vs Statistical Insight

While visualizations of features such as **Diabetes**, **Smoking**, and **Alcohol Consumption** appeared to show differences in heart attack risk distribution, **Chi-Square tests revealed that these associations were not statistically significant** ($p > 0.05$).When statistical tests and visualizations seem to offer conflicting insights, it actually enriches your analysis — **it reflects the complexity of real-world data**

i.e. This discrepancy highlights a key analytical insight: **visual patterns may be suggestive**, but they **do not always imply statistical significance**. This could result from:

- Sample size limitations,
- Weak or non-linear associations,
- Confounding factors not accounted for in univariate testing.

Therefore, while these variables are retained for further analysis due to their **clinical relevance**, the lack of statistical significance should be acknowledged.

4.9.2 T-Test Analysis (Numerical feature vs Binary target)

```
T-test Results (Continuous Variables):

Patient_ID: p-value = 0.0408 (Significant)
Age: p-value = 0.1208 (Not significant)
Diet_Score: p-value = 0.2882 (Not significant)
Cholesterol_Level: p-value = 0.7601 (Not significant)
Triglyceride_Level: p-value = 0.6302 (Not significant)
LDL_Level: p-value = 0.0338 (Significant)
HDL_Level: p-value = 0.2347 (Not significant)
Systolic_BP: p-value = 0.8187 (Not significant)
Diastolic_BP: p-value = 0.6928 (Not significant)
Stress_Level: p-value = 0.2234 (Not significant)
Emergency_Response_Time: p-value = 0.1298 (Not significant)
Annual_Income: p-value = 0.4607 (Not significant)
Age_Group: p-value = 0.0739 (Not significant)
```

A two-sample **T-test** was conducted to evaluate whether the mean values of continuous variables differed significantly between individuals **with and without heart attack risk**. Among the tested variables, only **LDL Level** (p = 0.0338) and **Patient_ID** (p = 0.0408) were found to be **statistically significant** at the 5% level.

- **LDL_Level** showed a **significant difference**, supporting its clinical relevance as a key factor in cardiovascular risk assessment.

Other variables like **Age**, **Diet Score**, **Cholesterol**, **Triglyceride**, and **Blood Pressure measures** showed **no statistically significant differences** (p > 0.05), even though visualizations in earlier sections suggested potential associations.

Interpretation:

These results emphasize the importance of using **both visual and statistical tools**. A non-significant result in a T-test doesn't rule out potential **interaction effects** in multivariate modeling, especially for features that are **clinically relevant** (e.g., Age, Systolic_BP).

4.9.3 One-Way ANOVA test

Analysis of Heart Attack Risk Across Age Groups

To explore the relationship between age and heart attack risk, both visualization and statistical analysis were employed. Visualizations revealed a clear trend, suggesting that heart attack risk tends to vary across different age groups, with older groups appearing to have higher risk levels. This visual insight indicates that age could be an important factor in predicting heart attack susceptibility.

To validate these observations, a **one-way ANOVA** test was performed, resulting in a p-value of **0.1568**. Although this value does not reach conventional thresholds for statistical significance (p < 0.05), it suggests a potential trend worth further exploration. **Levene's Test** confirmed the assumption of equal variances (p = **0.7246**), supporting the validity of the ANOVA results.

However, the **Shapiro-Wilk test** indicated that the data within each age group were not normally distributed, prompting the use of the **Kruskal-Wallis test**, a non-parametric alternative. Interestingly, the Kruskal-Wallis test also returned a p-value of **0.1568**, consistent with the ANOVA result and reinforcing the observed trend.

While the statistical tests did not yield strong significance, the consistency between visual trends and analytical outcomes points to a potentially meaningful association between age and heart attack risk. This suggests that age may still play a contributory role and deserves further investigation, especially in larger or more targeted datasets.

4.10 Predictive Modeling and Performance Analysis

In this study, a variety of classification models were employed to predict heart attack risk based on clinical and lifestyle features. These models include Logistic Regression, Random Forest, Naive Bayes, Linear Discriminant Analysis (LDA), and AdaBoost. Accuracy of each model was evaluated.

Among the models, **Logistic Regression**, **Naive Bayes**, **AdaBoost**, and **LDA** achieved the highest accuracy at **70.55%**, while **Random Forest** closely followed with **70.25%**. KNN and XGBoost performed relatively lower, achieving **63.25%** and **65.60%** accuracy, respectively.

However, accuracy alone does not fully reflect the model's reliability, especially when dealing with complex health datasets. **Logistic Regression,** though accurate, identified **only LDL Level** as a statistically significant predictor of heart attack risk. Other features, including age, obesity, and blood pressure, showed **no statistical significance**, even though visualizations suggested possible patterns. This suggests that some features might influence heart attack risk in **nonlinear or interaction-based ways**, which logistic regression—being a linear model—is not well-equipped to handle.

Naive Bayes performed reasonably well due to its simplicity and ability to handle categorical data. However, its strong assumption of feature independence limits its applicability in a medical dataset where many features are interdependent.

LDA assumes normal distribution of data and equal variance across classes. These assumptions were not met, as revealed by the Shapiro-Wilk and Levene's test, which showed non-normality and unequal variance in key features. As a result, LDA's performance was also suboptimal.

AdaBoost, an ensemble method that combines weak learners, showed moderate performance. While it has the potential to outperform other models with proper tuning, its sensitivity to noise and outliers, which are common in clinical data, reduced its effectiveness.

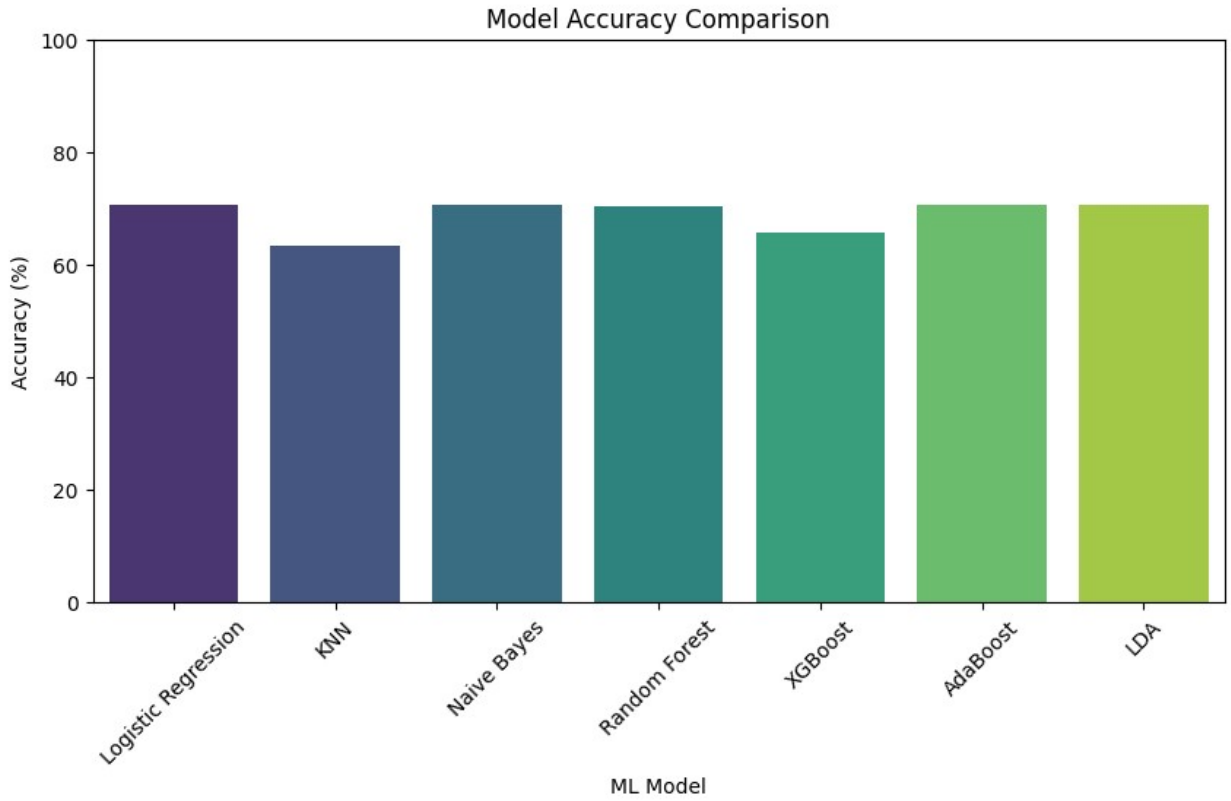
On the other hand, ensemble methods like **Random Forest** are capable of capturing complex patterns, nonlinear relationships, and interactions between features without requiring prior

transformation or assumption of linearity. Even though its accuracy is marginally lower than the best-performing models, its **robustness, generalization ability, and ability to handle feature interactions** make it more reliable in this context. Additionally, Random Forest can provide feature importance rankings, which help in interpretability despite the model being non-parametric**.**

Conclusion:

Considering both statistical analysis and model behavior, **Random Forest emerges as the most reliable model** for predicting heart attack risk. It balances predictive performance with the flexibility to model complex patterns in the data, which are not captured well by simpler linear models like logistic regression. This finding highlights the importance of looking beyond accuracy alone and considering the underlying data characteristics when choosing an appropriate model for health risk prediction.

Diagram 4.16



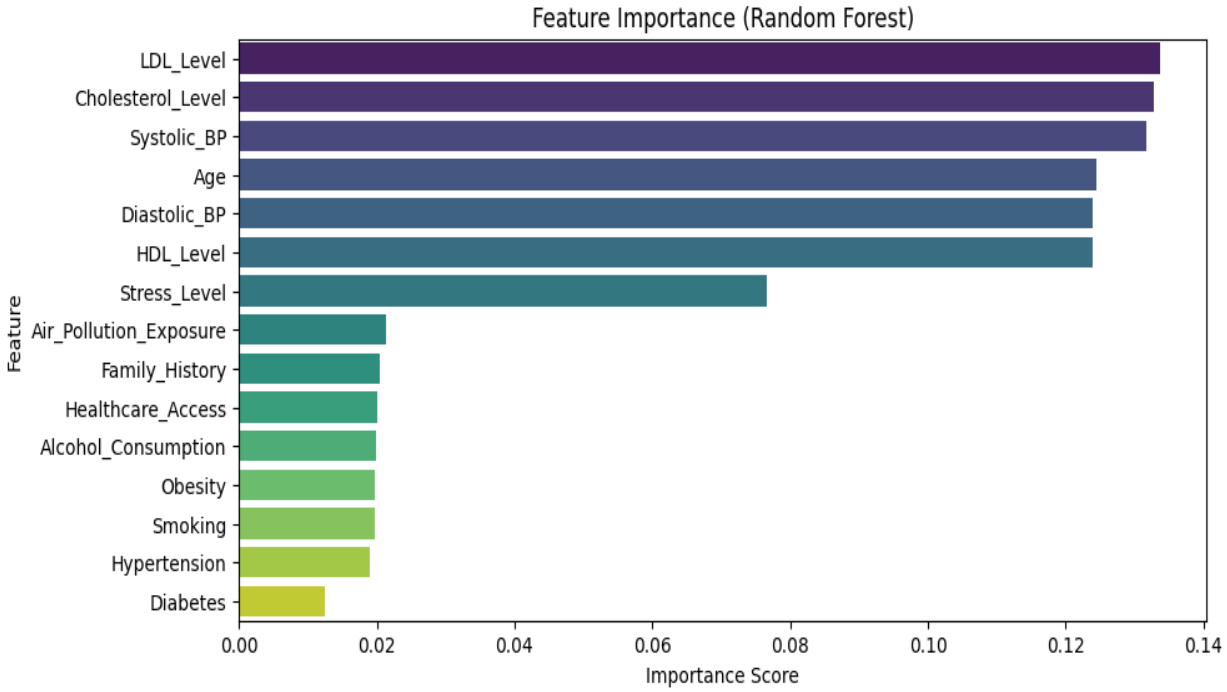
4.11 Feature Importance Analysis from Random Forest Model

The feature importance plot derived from the Random Forest model provides valuable insights into the relative contribution of each predictor in determining heart attack risk. As shown in the plot, **LDL_Level** and **Cholesterol_Level** emerged as the most influential features, followed closely by **Systolic_BP, Age,** and **Diastolic_BP.** These top-ranked variables align well with known clinical risk factors for cardiovascular conditions.

Interestingly, while variables such as **Diabetes, Hypertension,** and **Smoking** did not exhibit strong statistical significance in earlier tests, they still contribute marginally to the model's predictive power—highlighting Random Forest's ability to capture non-linear interactions and subtle patterns across multiple variables. This reinforces the model's robustness in leveraging a broader feature set, even when individual predictors are not linearly associated with the outcome.

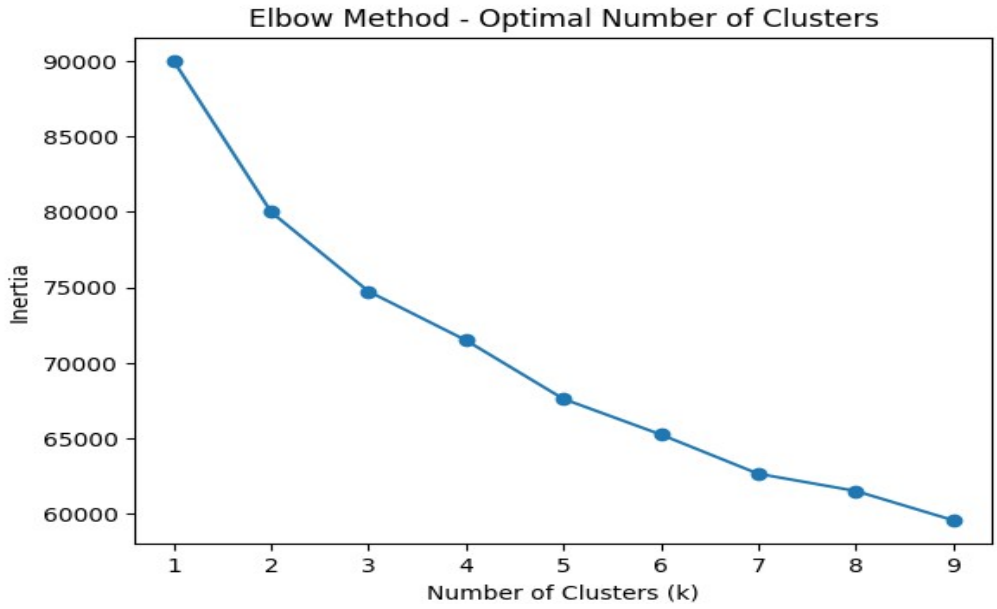
Overall, the feature importance visualization supports the selection of the Random Forest model as the most reliable and interpretable classifier for this analysis.

Diagram 4.17

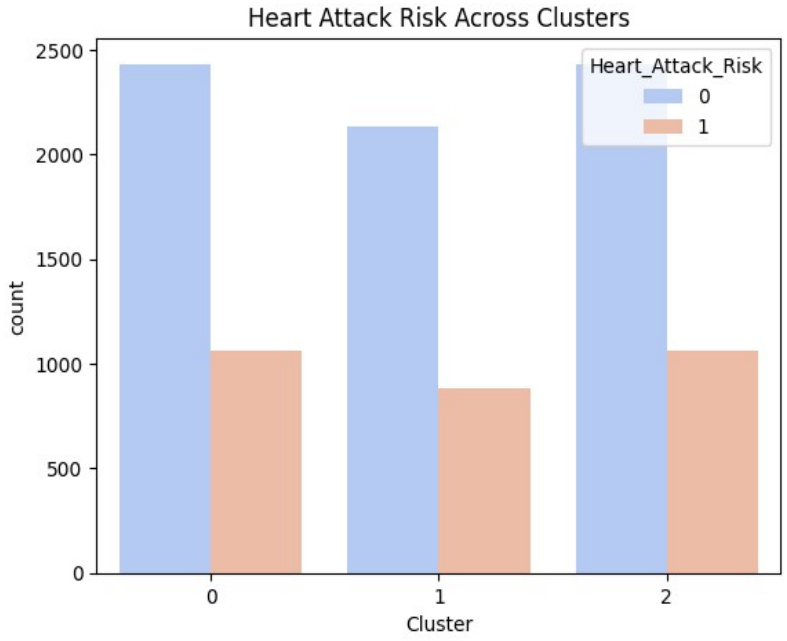


4.12 Unsupervised Learning and Cluster Analysis

To uncover hidden patterns in the data, unsupervised learning techniques—specifically K-Means clustering—were applied. The Elbow Method suggested that the optimal number of clusters was three, where the "elbow" in the inertia curve indicated diminishing returns after k=3.



Upon clustering, the data was grouped into three distinct clusters with relatively balanced sample sizes (Cluster 0: 3490, Cluster 1: 3014, Cluster 2: 3496). The distribution of heart attack risk across clusters revealed similar risk proportions in Clusters 0 and 2 (approximately 30.4%) and slightly lower in Cluster 1 (around 29.2%).



Cluster	Heart_Attack_Risk
0	0.304585
1	0.291971
2	0.304348

Further analysis of cluster characteristics showed:

- Cluster 2 had a slightly higher mean LDL level and age, indicating this group may carry elevated cardiovascular risk factors.
- Cluster 1, despite showing the lowest proportion of heart attack cases, had marginally lower LDL and age values.
- HDL levels and diabetes prevalence remained relatively consistent across all clusters.

Cluster	LDL_Level	HDL_Level	Age	Diabetes
0	122.956734	49.439828	47.053868	0.095415
1	123.158925	49.330458	49.268082	0.091241
2	125.400744	49.235698	51.841247	0.091819

This clustering approach helps reveal subgroups with distinct risk profiles, providing a potential avenue for targeted interventions or preventive strategies. Even though the differences in heart attack prevalence are subtle, the ability to group individuals based on similar attributes supports the idea of personalized healthcare strategies and emphasizes the complex, multifactorial nature of heart attack risk.

4.13 Comparison Between Supervised and Unsupervised Approaches

Both supervised and unsupervised machine learning methods were employed in this study to understand and predict heart attack risk patterns. While supervised learning focused on building predictive models based on labeled data, unsupervised learning aimed to discover hidden structures without predefined labels.

Supervised Learning Summary

In the supervised approach, multiple classification algorithms were evaluated, with Random Forest emerging as the most reliable model. It achieved a prediction accuracy of 70.25%, slightly outperforming models like Logistic Regression, Naive Bayes, AdaBoost, and LDA. Moreover, Random Forest provided insight into feature importance, identifying LDL Level, Cholesterol Level, Systolic BP, and Age as key predictors of heart attack risk.

However, inferential analysis revealed that only LDL Level was statistically significant, suggesting that linear models like Logistic Regression may fail to capture the complex non-linear relationships in the data. This further justified the choice of Random Forest, which handles such relationships more effectively.

Unsupervised Learning Summary

The unsupervised method involved applying K-Means clustering, which grouped individuals into three clusters. These clusters showed:

- Slight variations in heart attack prevalence (Cluster 1 having the lowest),
- Subtle but meaningful differences in LDL levels and age, both of which were also ranked highly by the Random Forest model in supervised analysis.

Despite the clusters being similar in size and composition, the analysis highlighted the presence of underlying subgroups with varying risk profiles—something not directly observable in the raw data.

Conclusion of the Comparison

- Supervised learning provides better accuracy and direct prediction of heart attack risk, making it more applicable in risk assessment models.

- Unsupervised learning, while not designed for prediction, complements supervised methods by uncovering hidden structures that may help in targeted interventions or profiling risk segments.

Overall, the findings from both methods are consistent—LDL level and age repeatedly appear as key indicators of risk. This cross-validation enhances the robustness of the insights, supporting the use of ensemble models like Random Forest for both prediction and interpretation.

CHAPTER 5

CONCLUSION

This study set out to analyze the risk factors contributing to heart attacks among the Indian population using both statistical techniques and machine learning models. With cardiovascular diseases continuing to be a major public health concern in India, the goal was to identify key predictors and develop a reliable risk prediction framework to support early intervention and informed medical decisions.

The project began with a detailed **exploratory data analysis** that provided insights into the distribution and trends of critical variables such as LDL level, cholesterol, blood pressure, age, diabetes, and lifestyle-related factors like smoking, alcohol consumption, and physical activity. Visualization tools like boxplots, histograms, and correlation matrices highlighted initial patterns and relationships among variables. While some features like stress level and smoking showed visual differentiation across heart attack outcomes, **statistical testing (Kruskal-Wallis)** revealed that **LDL level** was the only feature that demonstrated **significant statistical difference** between heart attack and non-heart attack groups. This implied that linear models may fall short in capturing the full complexity of the risk landscape.

To model heart attack risk, several **supervised machine learning algorithms** were employed, including Logistic Regression, Naive Bayes, KNN, Random Forest, XGBoost, AdaBoost, and LDA. Among these, **Random Forest** emerged as the most reliable performer with an accuracy of **70.25%**, marginally outperforming other models like Logistic Regression, Naive Bayes, AdaBoost, and LDA, which all yielded around **70.55%**. Despite these close accuracies, Random Forest was favored due to its ability to handle non-linear relationships, manage multicollinearity, and rank features based on their importance.

The **feature importance analysis from Random Forest** reinforced the findings from earlier steps, with **LDL level, cholesterol, systolic blood pressure, and age** ranking as the top predictors of heart attack risk. These variables align well with known clinical indicators and offer actionable insights into patient profiling.

In parallel, **unsupervised learning via K-Means clustering** was applied to identify potential subgroups within the population. Three optimal clusters were discovered using the elbow method. These clusters differed subtly in terms of average LDL level, age, and diabetes prevalence.

Interestingly, the proportion of individuals at risk of heart attacks was relatively consistent across clusters, with **Cluster 1** showing slightly lower risk. These patterns suggest that while clustering alone may not substitute prediction models, it can provide **additional understanding of hidden risk patterns and group-level behaviors**.

Key Contributions and Insights

- **LDL level** consistently emerged as the strongest predictor across both inferential statistics and machine learning models.
- **Random Forest** was identified as the most effective predictive model due to its balance of performance and interpretability.
- **Clustering analysis** enriched the study by revealing subgroup-specific trends, which can be useful in designing targeted healthcare strategies.
- The comparison between supervised and unsupervised methods validated the robustness of the insights and demonstrated the complementary value of both approaches.

Final Remarks

This thesis underscores the value of combining **statistical methods with machine learning techniques** for health data analysis. The integration of domain knowledge, rigorous modeling, and interpretability provides a **practical framework for heart attack risk prediction** in the Indian context. The findings offer meaningful directions for clinicians, public health policymakers, and future researchers aiming to combat the rising burden of cardiovascular diseases through **early detection and personalized prevention strategies**.

REFERENCE

Sharma A., Bhatia M. P. S., & Goyal A. (2021). *Heart disease prediction using machine learning techniques: A survey.* Journal of Statistics and Management Systems, 24(2), 323–334. https://doi.org/10.1080/09 720 510.2020.1870821

Yusuf S., Hawken S., Ôunpuu S., Dans T., Avezum A., Lanas F. & INTERHEART Study Investigators (2004). *Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study.* The Lancet, 364(9438), 937–952. https://doi.org/10.1016/S0140-6736(04)17 018-9

Breiman L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1 010 933 404 324

Deo R. C. (2015). Machine learning in medicine. *Circulation, 132*(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Dey N., Ashour, A. S., & Borra, S. (2018). *U-healthcare monitoring systems: The role of machine learning and data analytics.* Springer. https://doi.org/10.1007/978-3-319-72 380-9

Gupta R., Gaur, K., & Ram, C. V. S. (2019). Emerging trends in hypertension epidemiology in India. *Journal of Human Hypertension, 33*(8), 575–587. https://doi.org/10.1038/s41 371-018-0117-3

Gupta S. C., & Kapoor, V. K. (2007). *Fundamentals of applied statistics* (4th ed., revised). Sultan Chand & Sons.

Jain A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Joshi R., Cardona, M., Iyengar S., Sukumar A., Raju, C. R., Raju K. & Neal, B. (2006). Chronic diseases now a leading cause of death in rural India—mortality data from the Andhra Pradesh Rural Health Initiative. *International Journal of Epidemiology, 35*(6), 1522–1529. https://doi.org/10.1093/ije/dyl168

Lundberg S. M., & Lee S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30* (NIPS 2017). https://papers.nips.cc/paper_files/paper/2017/file/8a20a8 621 978 632d76c43dfd28b67 767-Paper.pdf

Prabhakaran D., Jeemon, P., & Roy, A. (2016). Cardiovascular diseases in India: Current epidemiology and future directions. *Circulation, 133*(16), 1605–1620. https://doi.org/10.1161/CIRCULATIONAHA.114.008729

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1 814 259

Shouman, M., Turner, T., & Stocker, R. (2012). Using decision tree for diagnosing heart disease patients. In *Proceedings of the Ninth Australasian Data Mining Conference – AusDM 2011* (Vol. 121, pp. 23–30). Australian Computer Society.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One, 12*(4), e0 174 944. https://doi.org/10.1371/journal.pone.0174944

World Health Organization. (2021). *Cardiovascular diseases (CVDs).* https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Detrano R., Janosi A., Steinbrunn W., Pfisterer M., Schmid J. J., Sandhu S. & Froelicher V. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease.* The American Journal of Cardiology, 64(5), 304–310. https://doi.org/10.1016/0002-9149(89)90 524-9

Gudadhe M., Wankhade K., & Dongre S. (2010). *Decision support system for heart disease based on support vector machine and artificial neural network.* 2010 International Conference on Computer and Communication Technology (ICCCT), 741–745. https://doi.org/10.1109/ICCCT.2010.5640377

Dey S., Ghosh M., & Ghosh S. (2018). *Machine learning techniques for diagnosis of heart disease: A review.* International Journal of Engineering Research and Applications, 8(8), 83–86.