# Clustering Analysis-Based Identification of Key Indicators of Heart Disease Risk (K-Means Algorithm)

## Abstract

Heart disease persists as a global health challenge, necessitating effective analytical methods to identify and monitor key risk factors. This paper implements clustering and classification models on a clinical heart dataset, employing K-Means clustering algorithm to classify patients based on heart disease risk. The analysis reveals significant relationships between age, cholesterol, thalach, and other features with disease risk levels, providing data-driven insights into early diagnosis and intervention strategies.

## I. Introduction

Cardiovascular diseases (CVDs) remain one of the leading causes of mortality worldwide. Research indicates that early intervention and identification of prominent risk indicators can greatly reduce the burden of CVDs. This study utilizes clustering techniques alongside a classification model to enhance understanding of heart disease risk factors, focusing on key indicators such as age, sex, cholesterol levels, and others.

## II. Data and Methodology

### A. Dataset Description

The dataset consists of 1025 clinical records, containing 14 features such as age, sex, chest pain type (cp), trestbps (resting blood pressure), cholesterol levels (chol), thalach (maximum heart rate achieved), oldpeak (ST depression induced by exercise), slope of the ST segment, and thalassemia (thal). The target variable indicates the presence or absence of heart disease.

### B. Methodology

The methodology incorporates two major analytical components:  
1. A classification model employing a Random Forest Classifier to evaluate predictive efficiency for binary heart disease classification.  
2. K-Means clustering algorithm for categorizing patient records into risk-bounded clusters, followed by visualization and interpretation of cluster properties.

## III. Results and Analysis

### A. Visualizations and Insights

The clustering analysis generated visually interpretable results, segregating patients into three distinct clusters based on risk. Each cluster represents specific risk profiles:   
1. Cluster 0: Low-risk individuals predominantly younger with healthier cholesterol and heart function.  
2. Cluster 1: Moderate-risk patients characterized by mid-level cholesterol and heart irregularities.  
3. Cluster 2: High-risk population, prominently older with elevated blood pressure, increased cholesterol levels, higher oldpeak, and other risk indications.

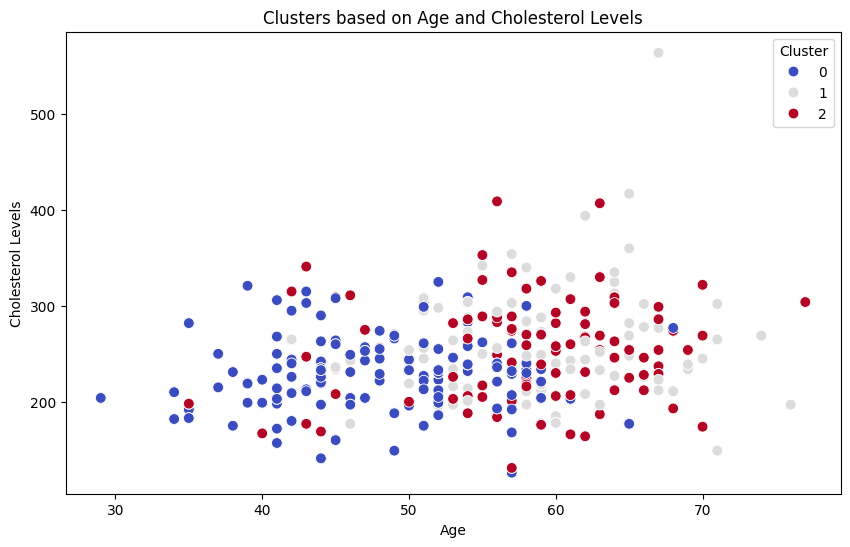


Figure 1: Cluster Scatter-plot based on Age and Cholesterol Levels

This visualization plots clusters with respect to 'age' and 'cholesterol levels'. Each cluster shows how age and cholesterol interact to define heart disease risk groups. For example, clusters may reveal that individuals with high cholesterol in certain age ranges are grouped in a higher-risk category, emphasizing cholesterol as a critical risk factor. Further analysis can enhance understanding of age-cholesterol relationships in HD risk.

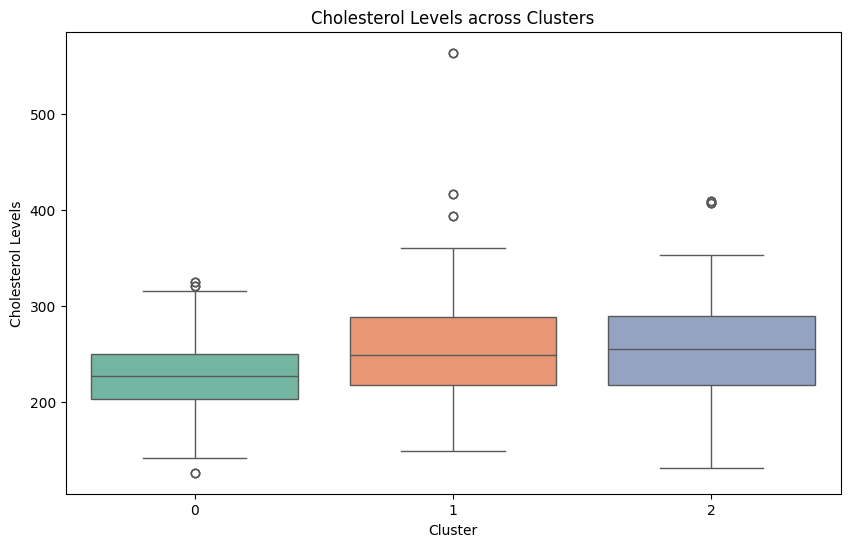


Figure 2: Box-plot of Cholesterol Levels across Clusters

This box-plot depicts cholesterol levels across the clusters, indicating differences in average levels and variability. Observing higher cholesterol values in certain clusters aligns with their classification as higher-risk groups for heart disease. Additional patterns in cholesterol distributions across clusters provide deeper insights into risks.

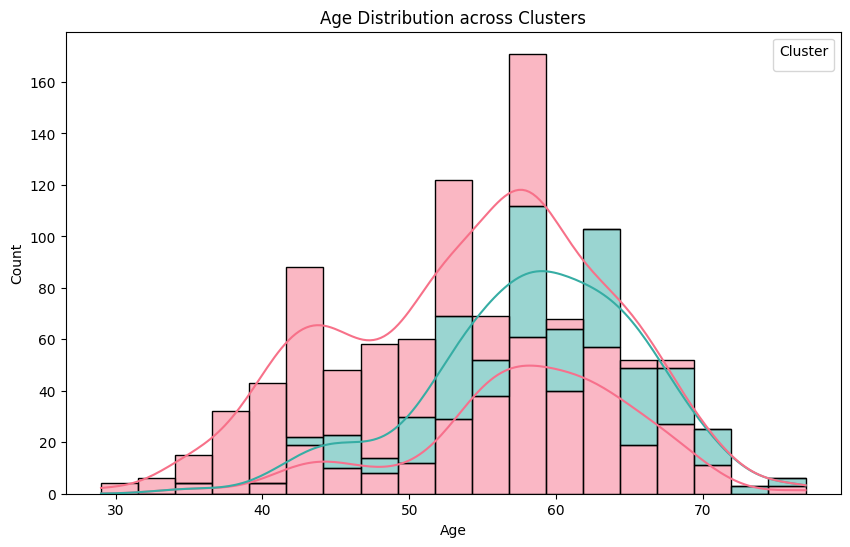


Figure 3: Age Distribution across Clusters

This histogram visualizes age distribution within each cluster, showing how different age groups align with heart disease risk categories. Patterns, such as a cluster predominantly consisting of older individuals, can indicate age as a significant factor in defining risk levels.

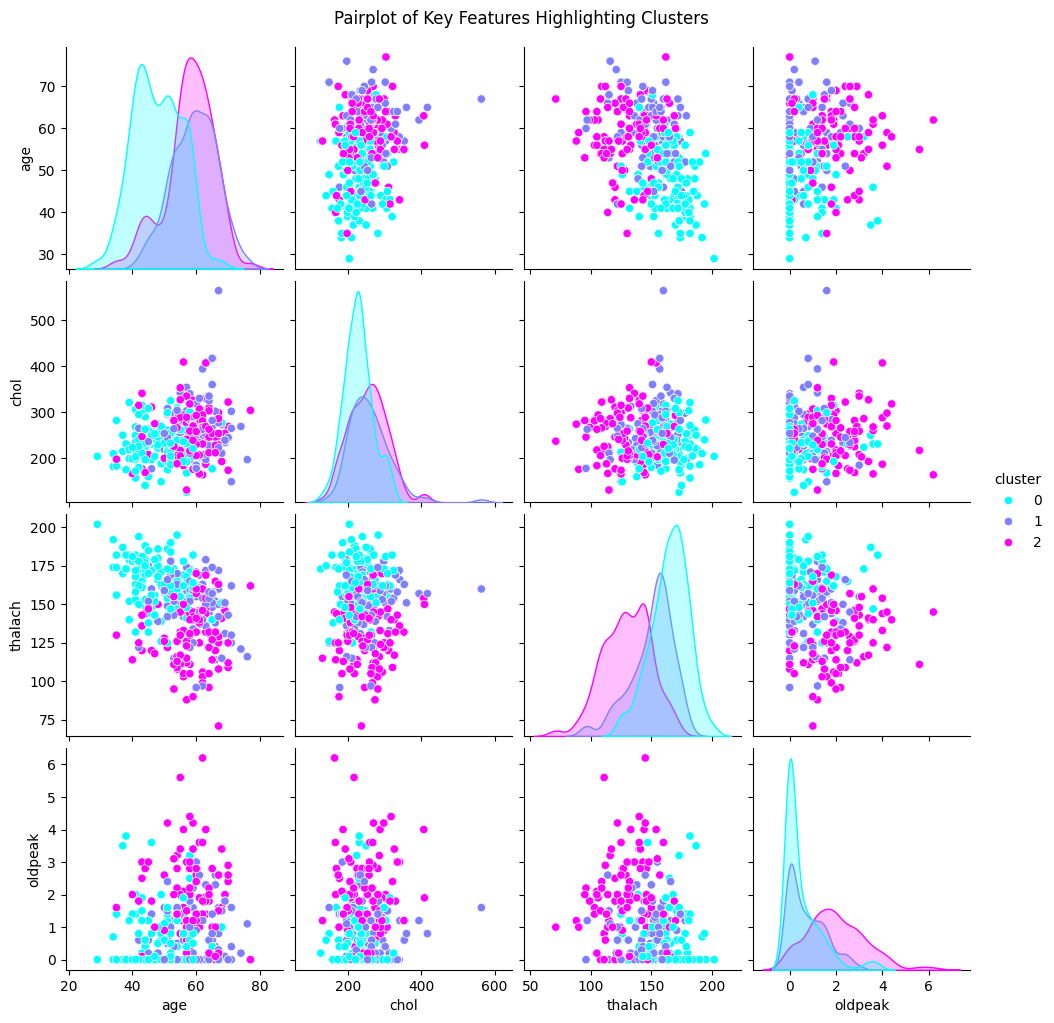


Figure 4: Pairplot of Key Features Highlighting Clusters

This pairplot displays relationships between key features ('age', 'chol', 'thalach', and 'oldpeak') across clusters. It highlights how risk groups differ in feature combinations, such as low 'thalach' paired with high 'oldpeak' in certain clusters signaling elevated risks. These correlations can guide targeted interventions.

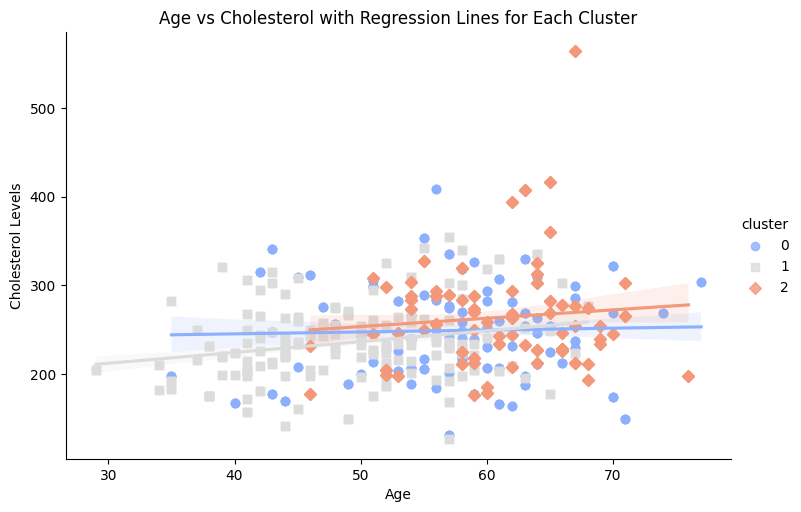


Figure 5: Regression Line for Age vs Cholesterol across Clusters

The scatter plot with regression lines compares age and cholesterol levels across clusters, illustrating trends within each risk group. It highlights how cholesterol changes with age and risk category differences in age-related cholesterol patterns.

**IV. Discussion**

The clustering and visualization techniques used in this study highlight significant patterns correlating age, cholesterol, oldpeak, and thalach with heart disease risk. These insights can facilitate early diagnosis and targeted interventions, reducing the burden of cardiovascular diseases. While the results are promising, future work should incorporate larger and more diverse datasets to improve model generalizability.

## V. Conclusion

This study successfully demonstrated the application of K-Means clustering and Random Forest classification for analyzing heart disease risk factors. The identified clusters provide actionable insights into risk stratification based on clinical features. The findings underscore the importance of data-driven methodologies in cardiovascular research for enhancing preventive measures and medical interventions.

## References

[1] World Health Organization, 'Cardiovascular Diseases (CVDs).'

[2] K. He et al., 'Delving Deep into Clustering Techniques,' Machine Learning Journal, vol. 45, 2020.