HEART DISEASE PREDICTION

Presented by
Kondapaneni Siva Rohit
Deepak Lingala
Uddav Ghimire

Date: 12/11/2024

INTRODUCTION

- A leading cause of global mortality, with millions of deaths annually.
- Early prediction can reduce risks and improve outcomes.
- This project uses machine learning in R Studio to predict heart disease, enabling data-driven medical decisions.

PROBLEM STATEMENT

- Heart disease is a leading global health challenge, contributing to high morbidity and mortality.
- **Challenges**
- Limited access to affordable healthcare and diagnostic tools.
- Identifying risk factors from patient data.
- Creating an interpretable and accurate predictive model.
- This project uses machine learning on public heart disease data to address these challenges.

OBJECTIVES

Primary Objective

- Develop a predictive model to detect heart disease using machine learning.

Specific Goals

- Identify key risk factors (e.g., age, cholesterol, blood pressure, lifestyle).
- Clean, preprocess, and explore data to uncover patterns.
- Implement and compare machine learning algorithms (e.g., decision trees, random forest).
- Derive insights to support early diagnosis and prevention.

DATA DESCRIPTION

Dataset Source

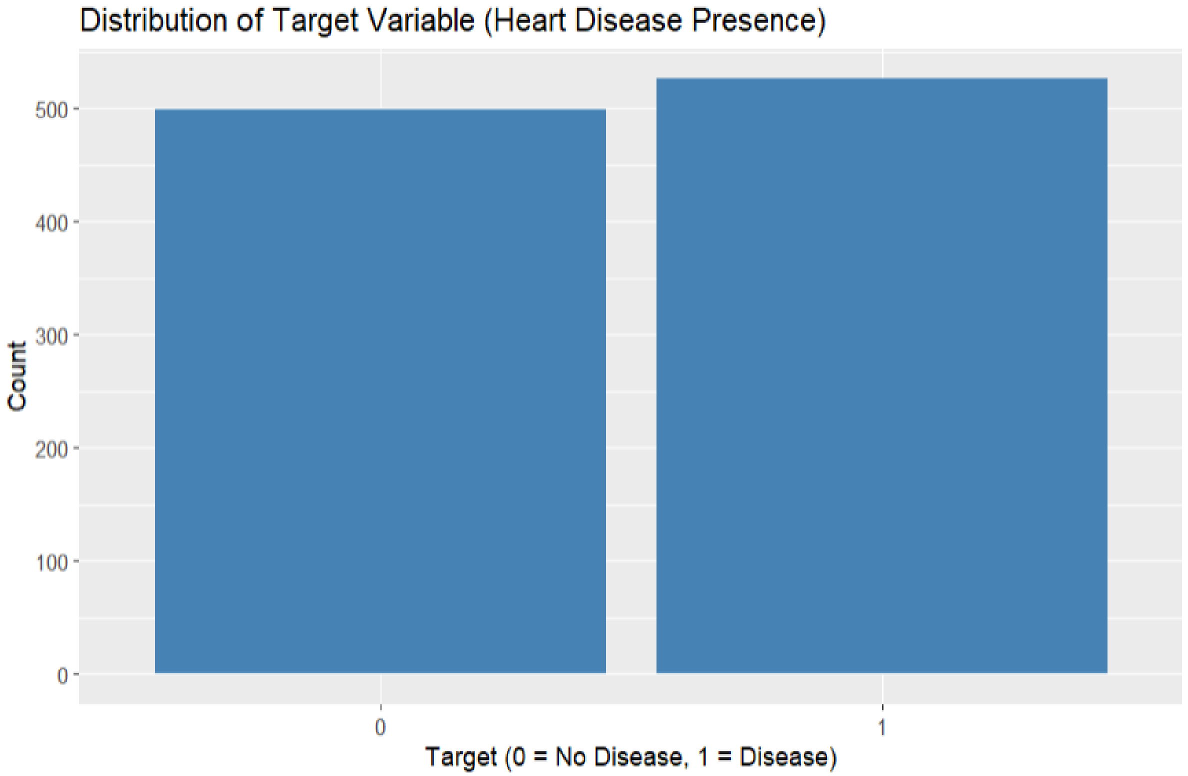
- Sourced from Kaggle: Heart Disease Dataset.
- **Overview**
- 1026 patient records with 14 features, including:
- Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol. Fasting Blood Sugar, Max Heart Rate, Exercise-Induced Angina, Oldpeak.
- Target: Heart disease presence (1) or absence (0).

Preprocessing

- Managed missing values and outliers.
- Standardized numerical variables.
- Encoded categorical variables for machine learning.

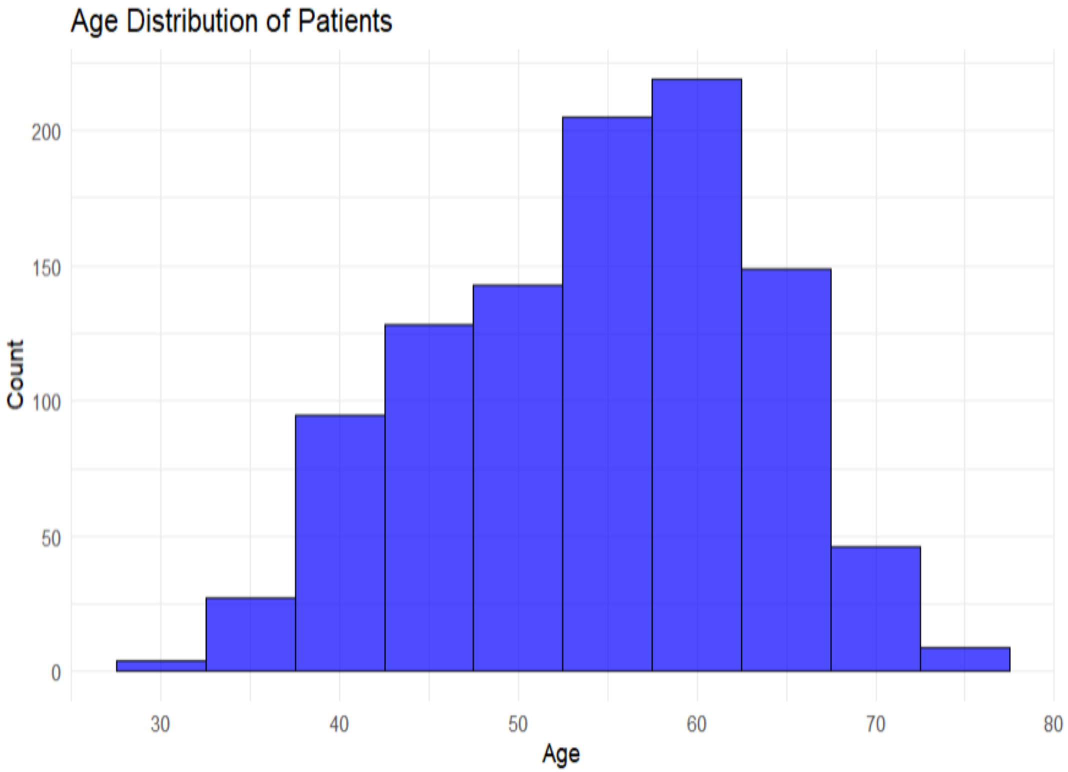
EXPLORATORY DATA ANALYSIS

- Balanced distribution between patients with and without heart disease
- There are slightly over 500 records for each category (0 and 1), suggesting a balanced dataset.



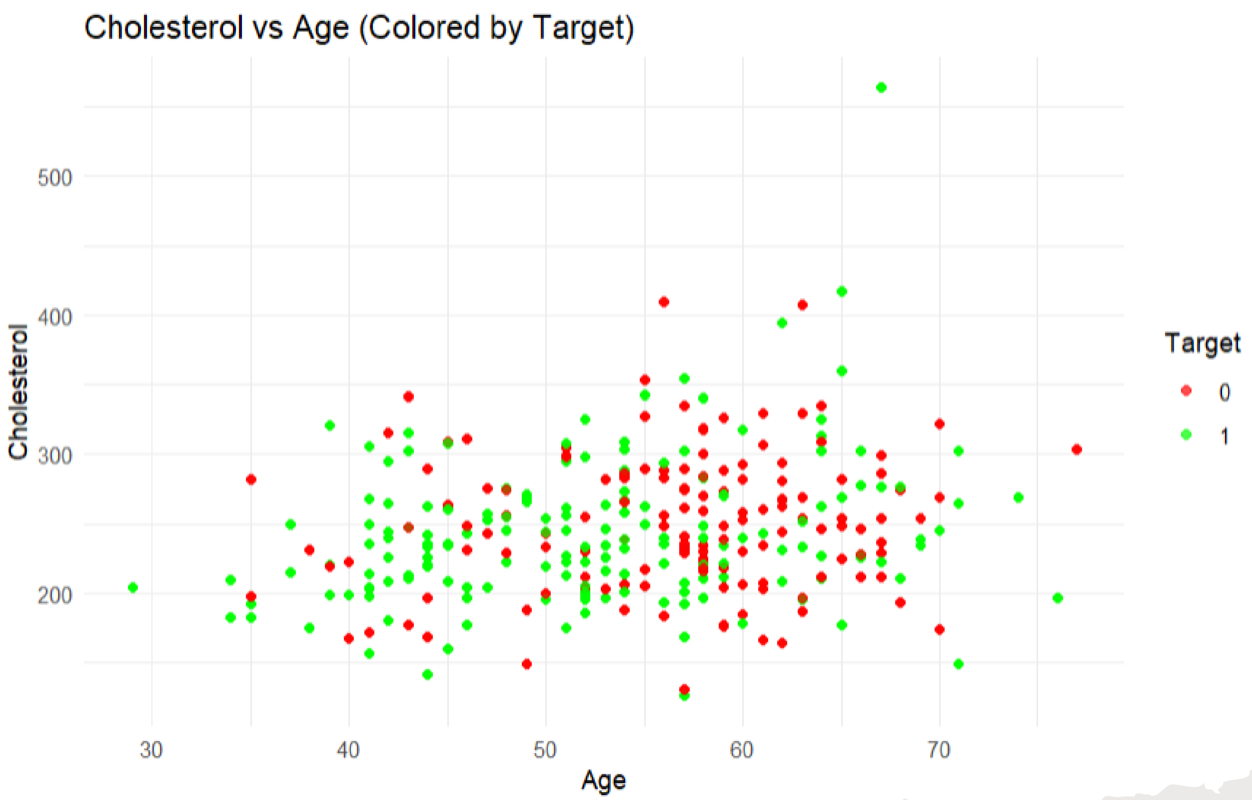
EXPLORATORY DATA ANALYSIS

- Majority of patients are between 40–65 years, with a peak around 55
- Most patients are within the 40-70 years range, with very few under 40 or over 70.
- This indicates that heart disease data is primarily concentrated in middle-aged to older adults, a demographic known to be more susceptible to cardiovascular conditions.



EXPLORATORY DATA ANALYSIS

- The scatterplot shows the relationship between cholesterol levels and age, with points colored by heart disease presence (0 = no disease, 1 = disease).
- No clear separation is visible between the two classes, indicating that cholesterol levels alone may not be sufficient to distinguish between individuals with and without heart disease.
- However, cholesterol levels are mostly concentrated between 200 and 400 mg/dL for both groups, with no significant trends correlating with age.



DATA MINING TECHNIQUES

- **Clustering Analysis:**
- Methodology: e.g., k-means, GMM Clustering.
- Key Findings: patterns in patient clusters.
- **Classification Models:**
- Algorithms used: Decision Trees, Random Forest
- Comparison of accuracy, precision, and other metrics.
- **Association Rules:**
- Associations between patient characteristics and heart disease.
- Apriori Algorithm and FP-Growth Analysis

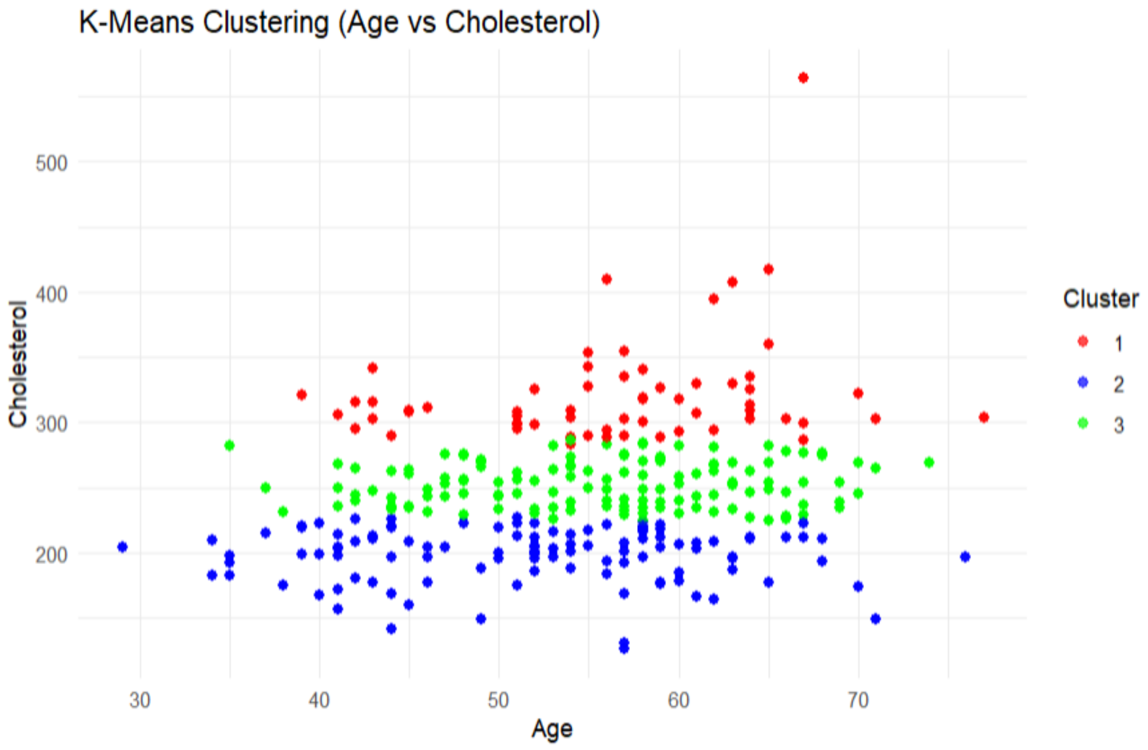
CLUSTERING ANALYSIS

K-Means Clustering Interpretation

- **Cluster 1 (Red)**: High cholesterol across all ages, indicating higher risk.
- **Cluster 2 (Green)**: Moderate cholesterol, mostly middle-aged (40–60 years).
- **Cluster 3 (Blue)**: Low cholesterol, predominantly younger (30–50 years), suggesting lower risk.

Insights

- Cholesterol is a stronger clustering factor than age.
- Adding features like heart rate or blood pressure may improve cluster insights.
- Supports subgroup identification for tailored healthcare interventions.



CLUSTERING ANALYSIS

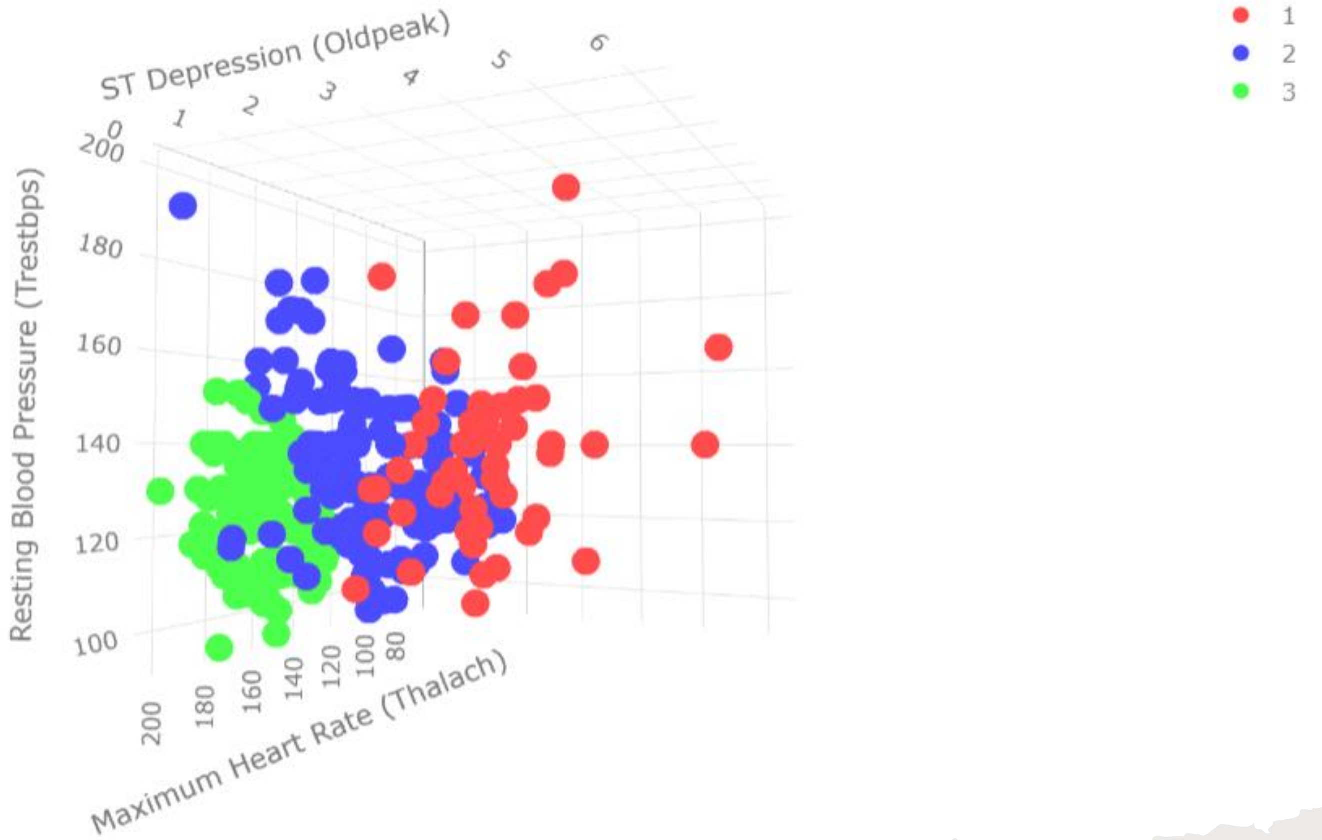
GMM Clustering Interpretation

Cluster Analysis

- **Cluster 1 (Red)**: High ST depression, moderate-high blood pressure – High-risk group.
- **Cluster 2 (Blue)**: Moderate ST depression, slightly lower heart rate and blood pressure – Medium risk.
- **Cluster 3 (Green)**: Low ST depression, high heart rate, low blood pressure – Low-risk group.

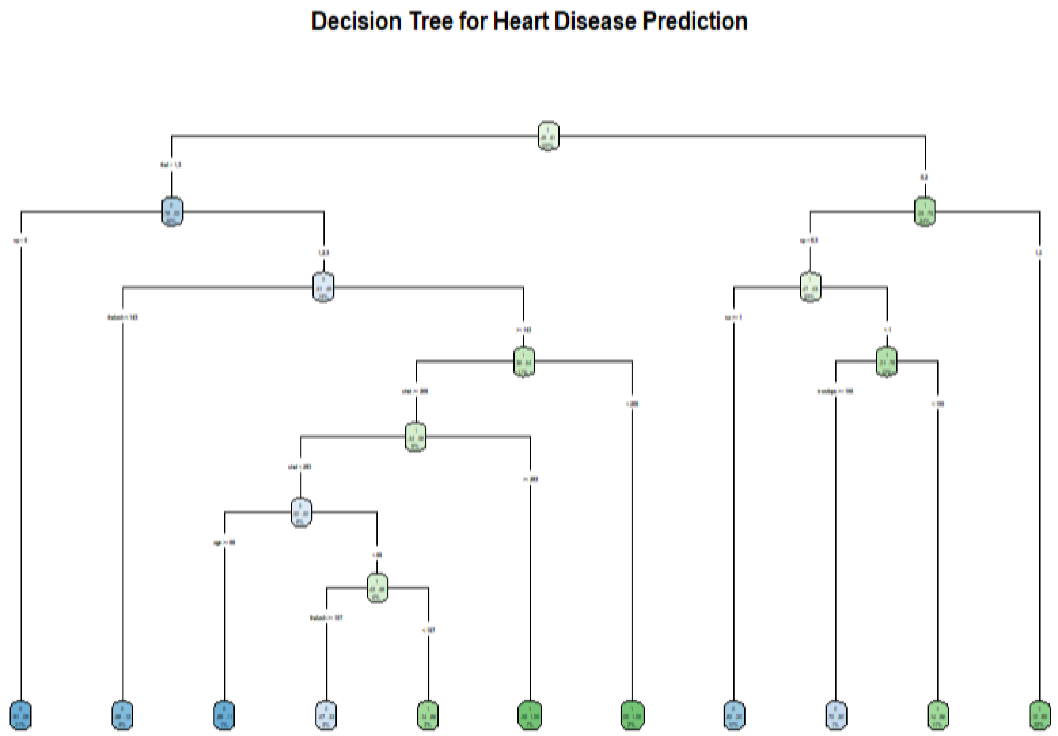
Insights

- Cluster 1: High-risk; needs closer monitoring.
- Cluster 2: Medium risk; signs of cardiac stress.
- Cluster 3: Low-risk; stable cardiac performance.



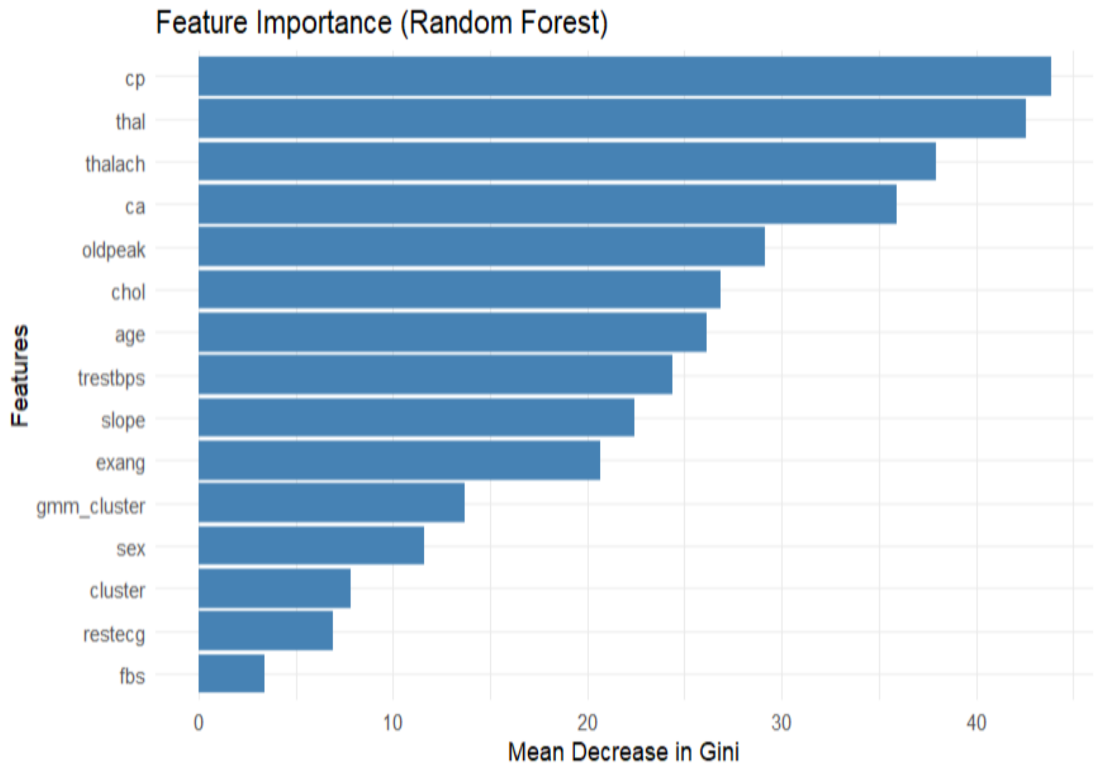
CLASSIFICATION MODELS

- **Decision Tree Interpretation**
- **Key Features**
- `thal` (Thalassemia levels) is the most influential predictor of heart disease.
- `op` (Oldpeak): Higher values indicate greater cardiac stress.
- `thalach` (Max Heart Rate): Lower values increase the likelihood of heart disease.
- Clustering features like `dbscan_cluster` and `cluster_new` further refine predictions.
- **Insights**
- Key predictors (`thal`, `op`, `thalach`) provide clear thresholds for risk assessment.
- Higher `op` combined with lower `thalach` values is strongly associated with heart disease.
- Clustering features improve decision-making in complex cases.



CLASSIFICATION MODELS

- **Random Forest Interpretation**
- **Important Features**:
  `thal` (Thalassemia): Strongest predictor.
  `cp` (Chest Pain Type): Significant in risk assessment.
  `ca` (Major Vessels Colored by Fluoroscopy): Strongly linked to heart disease.
- **Moderately Important Features**:
  `thalach` (Max Heart Rate) and `age`.
- **Least Important Features**:
  `fbs` (Fasting Blood Sugar) and `hc_cluster_selected` contribute minimally and may be excluded in future models.



ASSOCIATION RULES

- Key Insights:
- Both algorithms highlight sex, exercise-induced angina (exang), slope, thalassemia (thal), and GMM cluster assignments as significant features.
- Apriori:
- Strongest rule has a lift of 2.89 and perfect confidence (100%).
- FP-Growth:
- Supports up to 85.07% of transactions with high lift 2.89 and confidence.
- Top Rule Example (Both Methods):
- {sex=1, exang=0, slope=2, thal=2, gmm_cluster=3} => oldpeak=[0, 0.1)
- Male patients without exercise-induced angina and specific attributes are strongly associated with low ST depression.

RESULTS AND DISCUSSION

- **Key Findings**
- **Random Forest**: Best model with 100 % accuracy, 100 % sensitivity, and 100 % F1 Score.
- **Decision Trees**: Interpretable with 85 % accuracy.
- **Feature Importance**
- Key Predictors: `cp` (Chest Pain), `thalach` (Max Heart Rate), `oldpeak` (ST Depression), `trestbps` (Resting BP).
- **Insights**
- Higher cholesterol and ST depression levels increase heart disease risk.
- Random Forest offers high accuracy but less interpretability, while Decision Trees are easier to interpret with slightly lower accuracy.
- **Implications**
- Supports early detection and personalized treatment for heart disease.

CHALLENGES AND LIMITATIONS

Challenges

- **Data Quality**: Missing values (cholesterol, blood pressure) and outliers affected stability.
- **Dataset Size**: Only 1026 records, limiting generalizability.
- **Feature Correlation**: High correlation (e.g., age and cholesterol) complicated predictor selection.
- **Model Trade-Offs**: Balancing interpretability (Decision Trees) and performance (Random Forest).

Limitations

- **Dataset Scope**: Limited demographics; lacks diverse population representation.
- **Simplified Features**: Excludes factors like family history or lifestyle habits.
- **Model Complexity**: Advanced models (Random Forest) are less interpretable.

CONCLUSION

- **Summary of Findings**
- Random Forest achieved accuracy 1 in predicting heart disease.
- Key predictors: chest pain type, max heart rate, ST depression, resting BP.
- Data-driven approaches show promise for early detection and intervention.
- **Key Takeaways**
- Predictive models can efficiently identify at-risk patients.
- Insights support targeted prevention and personalized treatments.
- Interpretable models like Decision Trees remain valuable in clinical use.
- **Future Scope**
- Expand datasets to include diverse populations.
- Add clinical and lifestyle features for comprehensive analysis.
- Use explainable AI to enhance model interpretability in clinical settings.
- **Final Thoughts**
- Combining healthcare data with machine learning offers impactful solutions for preventive healthcare and decision-making.

THANK YOU

CONTACT INFORMATION

- Kondapaneni Siva Rohit
- sivarohit2002@yahoo.com
- •Deepak Lingala
- •deepaklingala@arizona.edu
- •Uddav Ghimire
- •uddav@arizona.edu