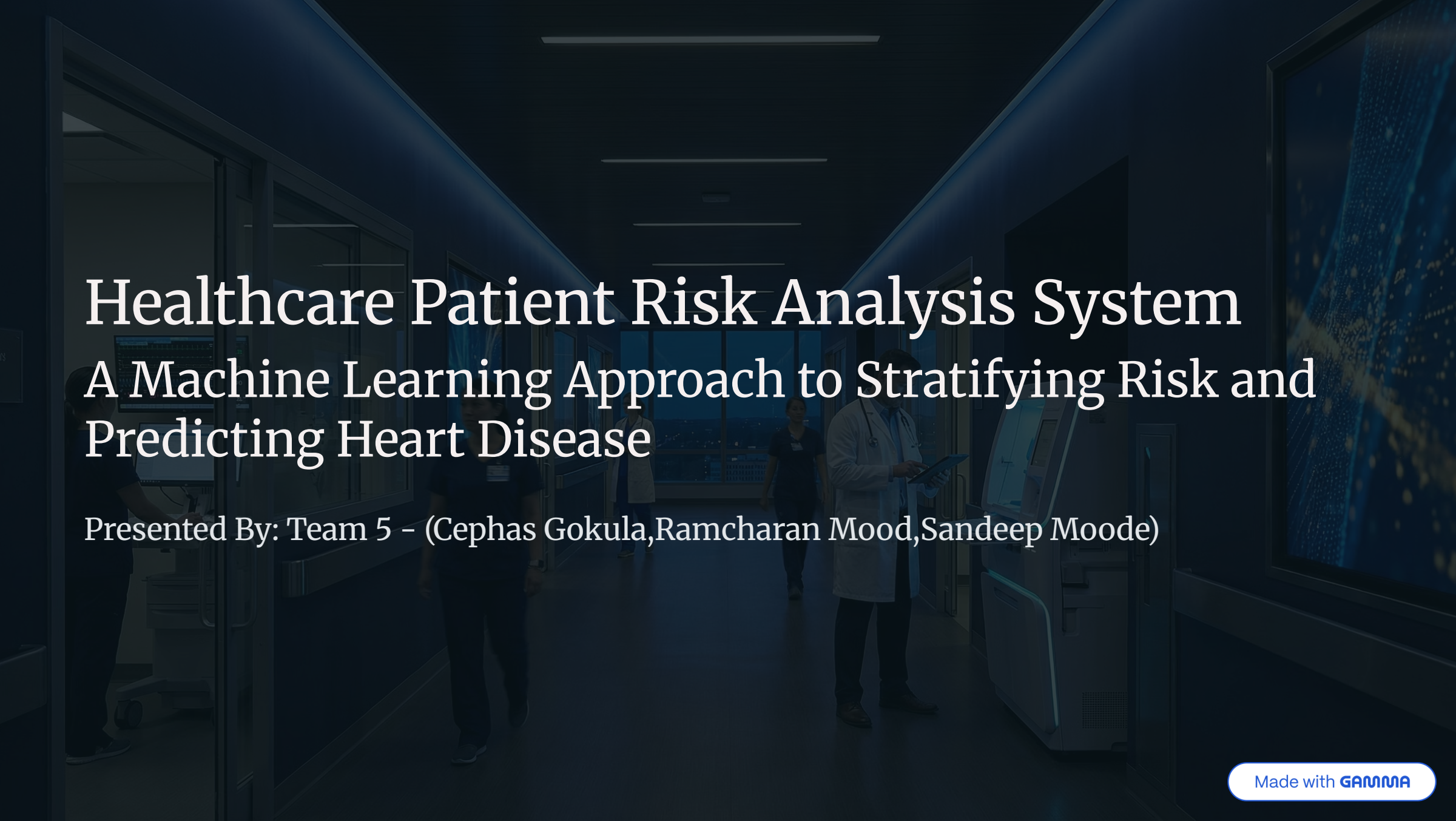


Healthcare Patient Risk Analysis System

A Machine Learning Approach to Stratifying Risk and Predicting Heart Disease

Presented By: Team 5 - (Cephas Gokula,Ramcharan Mood,Sandeep Moode)

The Challenge

Problem Statement

The Challenge

Cardiovascular diseases are the leading cause of death globally. Early detection is often missed due to the complexity of analyzing multiple patient risk factors simultaneously.

Current Gap

Traditional diagnosis relies heavily on manual symptom checks, which can miss hidden patterns or "at-risk" groups that don't show classic symptoms.

The Goal

Build an automated pipeline that not only **predicts** disease but also **groups** patients by similarity and **removes** data anomalies to improve accuracy.

Dataset Overview

Source & Volume

- **Source:** UCI Heart Disease Dataset (Cleveland Database).
- **Volume:** ~303 Patient records.

Target Variable

Diagnosis (0 = No Disease, 1 = Disease Present).

Key Attributes (14 features)

- **Demographics:** Age, Sex.
- **Vitals:** Resting Blood Pressure (trestbps), Cholesterol (chol), Max Heart Rate (thalach).
- **Clinical:** Chest Pain Type (cp), Fasting Blood Sugar, ST Depression (oldpeak).

Methodology – The 3-Stage Pipeline

Stage 1: Outlier Detection (Unsupervised)

Cleaning the data by removing impossible or extreme biological values.

Stage 2: Clustering (Unsupervised)

Grouping patients into "profiles" to understand risk demographics.

Stage 3: Classification (Supervised)

Training a model to predict the probability of heart disease for new patients.

Technique 1 – Outlier Detection

Algorithm: Isolation Forest

Why? Medical data often contains errors (e.g., Blood Pressure entered as 0 or 300) or rare genetic anomalies that confuse predictive models.

How it works:

- Isolates observations by randomly selecting a feature and a split value.
- Anomalies are "few and different," so they are isolated quickly (fewer splits).

Outcome: Identified and flagged ~5% of patients as outliers (extreme Cholesterol/BP) to ensure robust training.

Technique 2 – Patient Clustering

Algorithm: K-Means Clustering

01

Remove Diagnosis Label

Removed the diagnosis label (unsupervised).

02

Group by Similarity

Grouped patients based purely on physiological similarity.

03

Visualize with PCA

Used **PCA (Principal Component Analysis)** to visualize these groups in 2D.

> **Key Insight:** Discovered distinct patient profiles, e.g., "Cluster A" (Young, High Max Heart Rate, Low Risk) vs. "Cluster B" (Elderly, High Cholesterol, High Risk).

Technique 3 – Classification (Prediction)

Algorithm: Random Forest Classifier

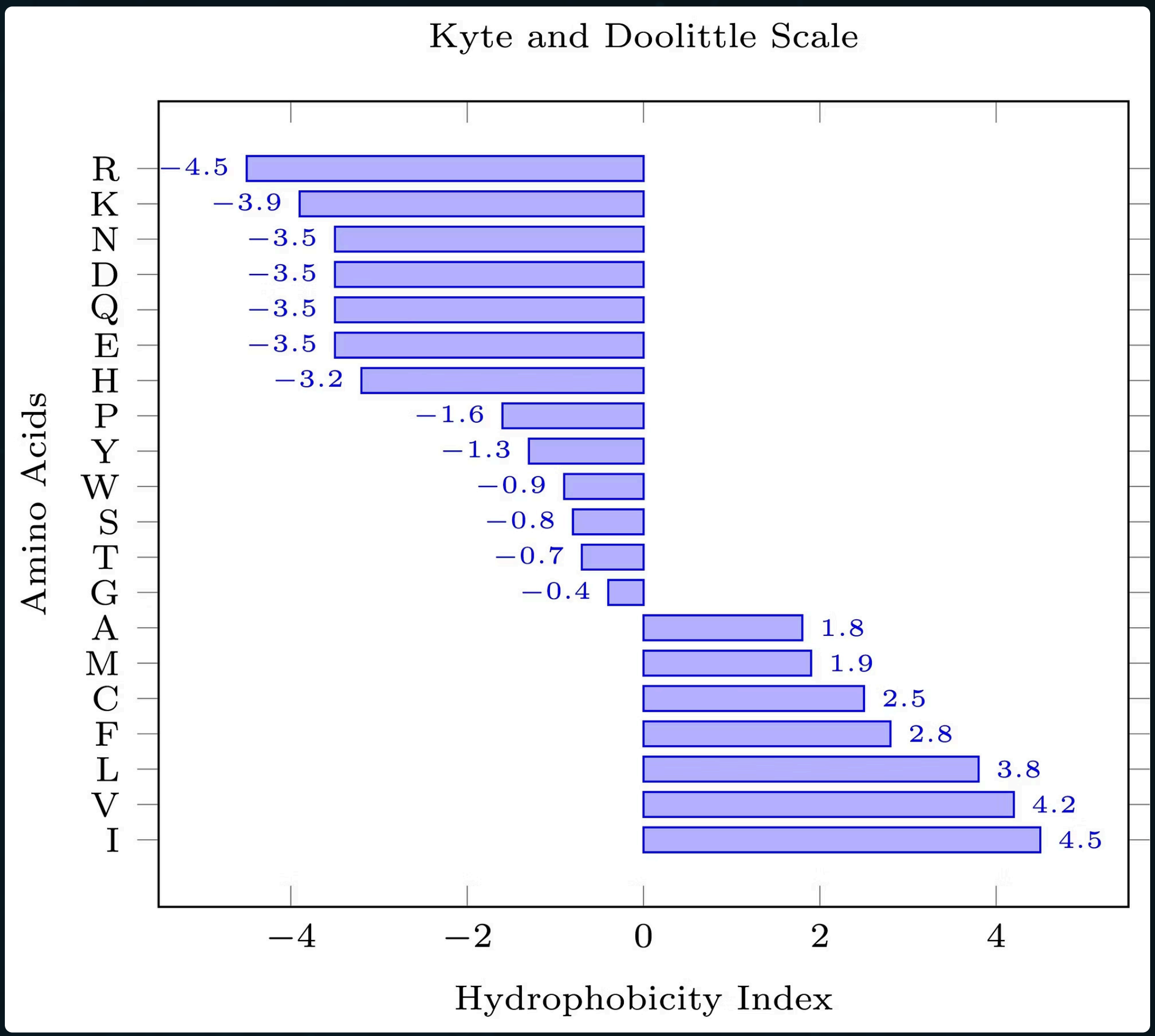
Why? It handles non-linear relationships well (e.g., the risk of Age combined with Cholesterol isn't a straight line) and provides **Feature Importance**.

Training: 80% Training Data, 20% Testing Data.

Metric Focus: Recall (Sensitivity)

Why Recall? In healthcare, a False Negative (telling a sick person they are healthy) is life-threatening. We prioritized catching *all* cases.

Results & Performance



Model Accuracy

~85-88% (or whatever your specific code output).

Confusion Matrix

- True Positives: Correctly identified disease.
- False Negatives: Missed diagnoses (minimized).

Feature Importance Chart

Top Predictors: Chest Pain Type (`cp`), Thalassemia, and Oldpeak (ST depression) were found to be more significant predictors than just Cholesterol levels.

Clinical Implications

Risk Stratification

Hospitals can use the "Clusters" to design standard care paths for specific patient groups.

Triage Support

The Classification model can serve as a "First Opinion" tool in emergency rooms to prioritize high-risk patients.

Data Quality

The Outlier Detection module acts as an automated quality control for Electronic Health Records (EHR).

Conclusion & Future Scope

Conclusion

The project successfully demonstrated that combining unsupervised learning (grouping) with supervised learning (predicting) yields richer insights than prediction alone.

Future Work

- Integrate Deep Learning (Neural Networks) for larger datasets.
- Deploy as a Web App (Streamlit/Flask) for doctors to input vitals and get real-time risk scores.

Q&A