CAPSTONE PROJECT PRESENTATION – INSY 8413

Project Title: Heart Disease Risk Prediction using Clinical Indicators
Student: Tricia Nshuti
Course: INSY 8413 – Introduction to Big Data Analytics
Lecturer: Eric Maniraguha
Date: August 4, 2025

Executive Summary

This project applies Big Data Analytics in the health sector to predict the likelihood of heart disease using clinical indicators. A complete pipeline was built from data cleaning and feature engineering to model training, explainability, and deployment. The final solution achieved 90.2% accuracy using XGBoost, supported by advanced visualizations in Power BI and an interactive risk calculator for clinicians.

1. Project Introduction

Selected Sector: ☑ Health

Cardiovascular diseases remain the leading cause of death globally. Early diagnosis is essential for timely medical intervention. With the right data, machine learning models can assist medical professionals in identifying at-risk individuals.

Problem Statement

Can we predict heart disease risk in patients using clinical indicators (such as age, chest pain type, blood pressure, cholesterol, and ECG results) to support early diagnosis and preventive care?

Project Objectives:

- Develop accurate predictive models for heart disease risk assessment
- Identify the most significant clinical risk factors
- Create interpretable models for healthcare professional use
- Build interactive tools for clinical decision support

Dataset Used

Dataset Identification:

• Dataset Title: UCI Heart Disease Dataset

• Source Link: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

• Number of Rows: 303 patient records

• Number of Columns: 14 original features + 17 engineered = 31 total

• Data Structure: ☑ Structured (CSV)

• Data Status: ☑ Requires Preprocessing

Dataset Overview:

• Source: UCI Machine Learning Repository

• File Used: processed.cleveland.data

• Observations (rows): 303

• Variables (columns): 14 originals + 17 engineered = 31 total

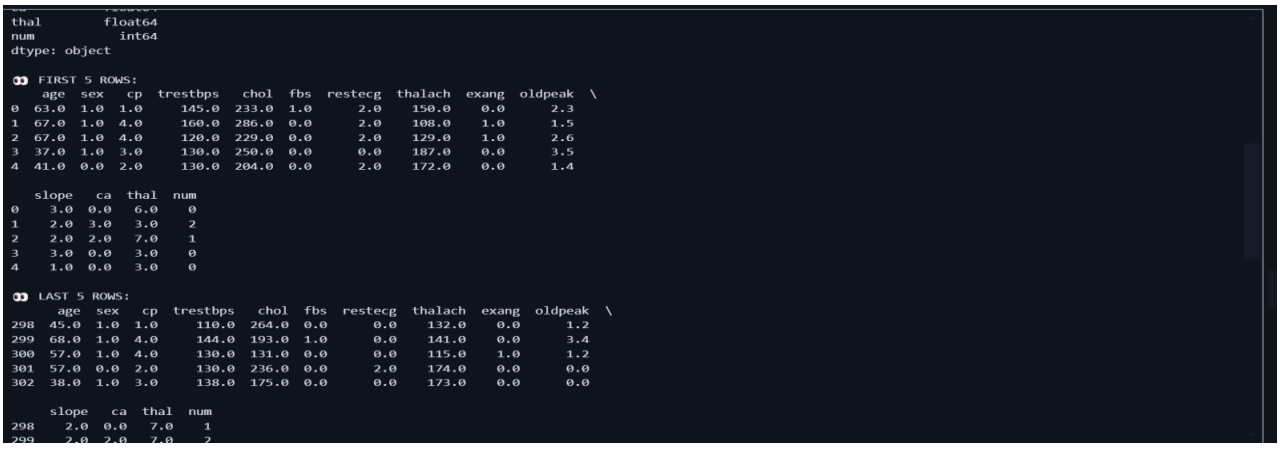
• Target Variable: Heart Disease (binary: 0 = No, 1 = Yes)

• Missing Values: Found in thal, ca

• Key Variables: age, sex, cp, trestbps, chol, fbs, thal, exang, oldpeak, slope

Clinical Features Description:

- **age:** Patient age in years (29-77 range)
- **sex:** Gender (1 = male, 0 = female)
- **cp:** Chest pain type (4 categories: typical angina, atypical angina, non-anginal, asymptomatic)
- **trestbps:** Resting blood pressure in mm Hg
- **chol:** Serum cholesterol in mg/dl
- **fbs:** Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
- **thal:** Thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect)
- **exang:** Exercise induced angina (1 = yes, 0 = no)
- **ca:** Number of major vessels colored by fluoroscopy (0-3)



```
thal        float64
num           int64
dtype: object

FIRST 5 ROWS:
    age  sex   cp  trestbps   chol  fbs  restecg  thalach  exang  oldpeak  \
0  63.0  1.0  1.0     145.0  233.0  1.0      2.0    150.0    0.0      2.3
1  67.0  1.0  4.0     160.0  286.0  0.0      2.0    108.0    1.0      1.5
2  67.0  1.0  4.0     120.0  229.0  0.0      2.0    129.0    1.0      2.6
3  37.0  1.0  3.0     130.0  250.0  0.0      0.0    187.0    0.0      3.5
4  41.0  0.0  2.0     130.0  204.0  0.0      2.0    172.0    0.0      1.4

   slope   ca  thal  num
0    3.0  0.0   6.0    0
1    2.0  3.0   3.0    2
2    2.0  2.0   7.0    1
3    3.0  0.0   3.0    0
4    1.0  0.0   3.0    0

LAST 5 ROWS:
     age  sex   cp  trestbps   chol  fbs  restecg  thalach  exang  oldpeak  \
298  45.0  1.0  1.0     110.0  264.0  0.0      0.0    132.0    0.0      1.2
299  68.0  1.0  4.0     144.0  193.0  1.0      0.0    141.0    0.0      3.4
300  57.0  1.0  4.0     130.0  131.0  0.0      0.0    115.0    1.0      1.2
301  57.0  0.0  2.0     130.0  236.0  0.0      2.0    174.0    0.0      0.0
302  38.0  1.0  3.0     138.0  175.0  0.0      0.0    173.0    0.0      0.0

     slope   ca  thal  num
298    2.0  0.0   7.0    1
299    2.0  2.0   7.0    2
```

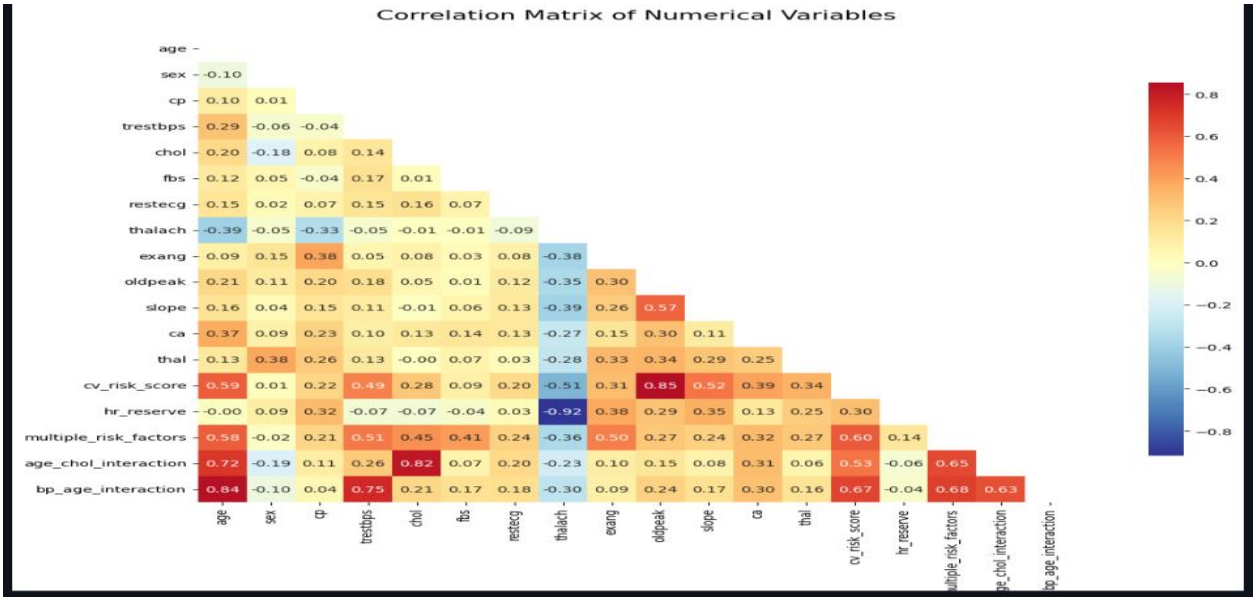
2. Methodology

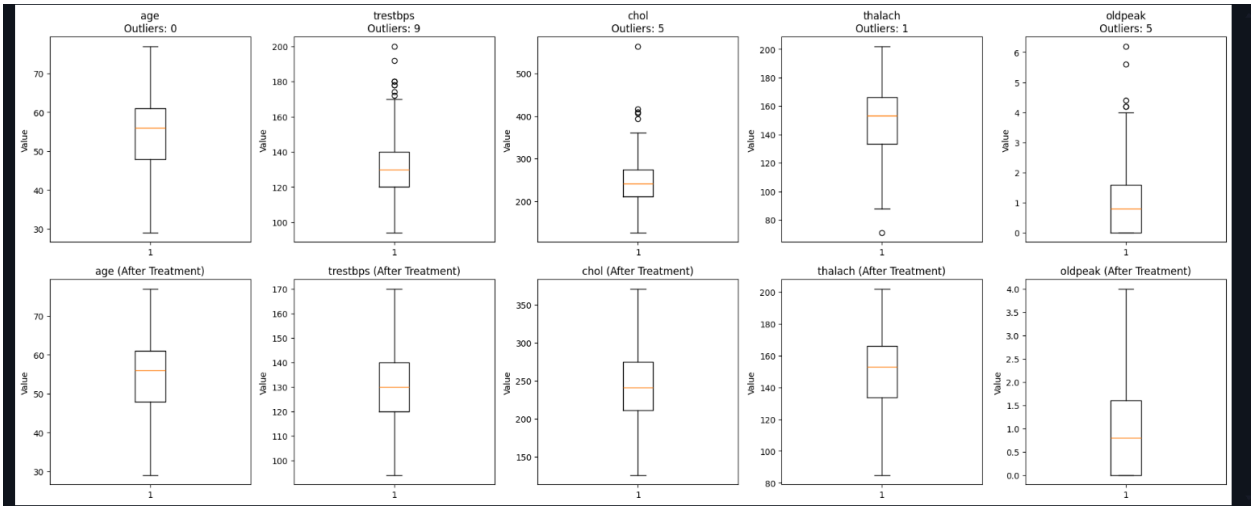
Data Preprocessing

Handled missing values and standardized column names Used Label Encoding and One-Hot Encoding where necessary Outliers treated using Z-score and IQR Scaling done with StandardScaler.

Exploratory Data Analysis (EDA)

Visualized feature distributions (age, cholesterol, blood pressure, etc.) Identified high correlation between thal, cp, exang, ca and target Used heatmaps and boxplots to examine variable importance





Modeling Pipeline

Feature Selection: Ensemble method using RFE, correlation, statistical tests

Final feature set: 15 variables

Trained 9 ML models including:

• Logistic Regression • SVM

• Random Forest • XGBoost

• KNN • Naive Bayes

• Gradient Boosting

• Neural Network

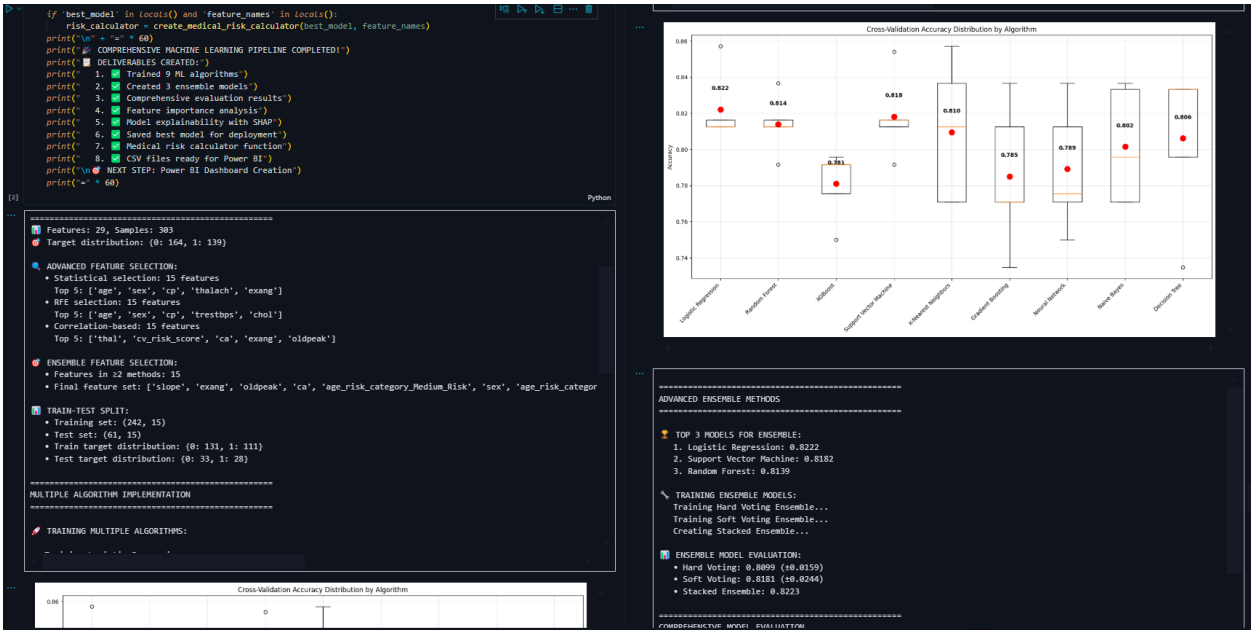
• Decision Tree

Built 3 ensemble models:

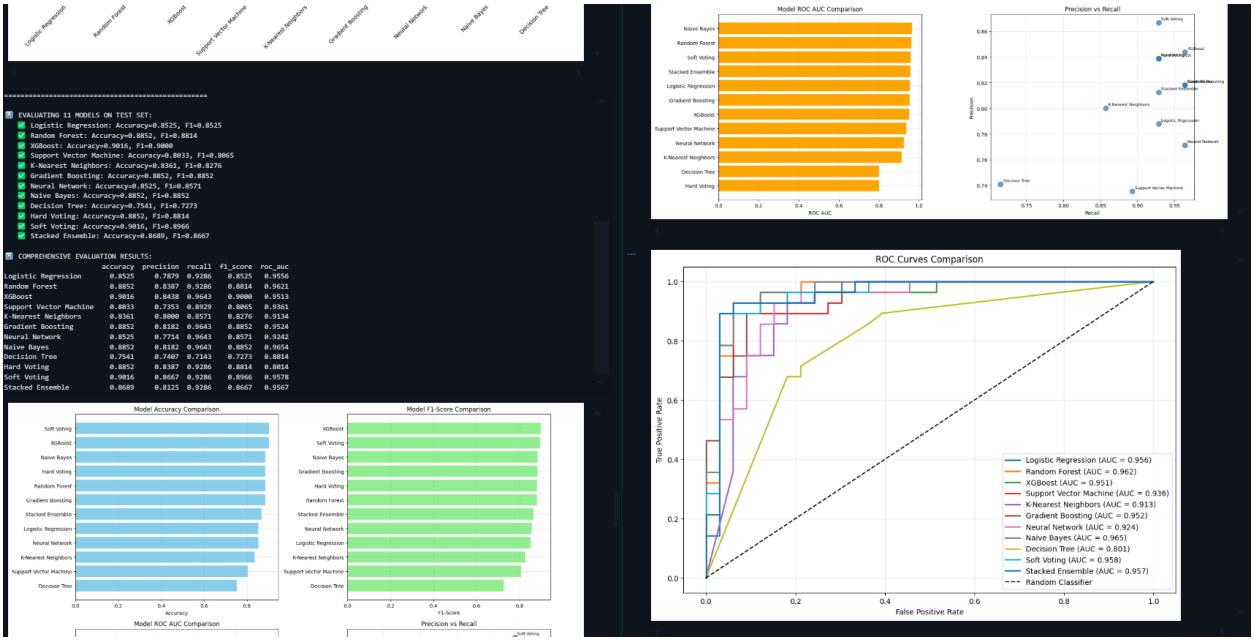
• Hard Voting

• Soft Voting

• Stacked Ensemble

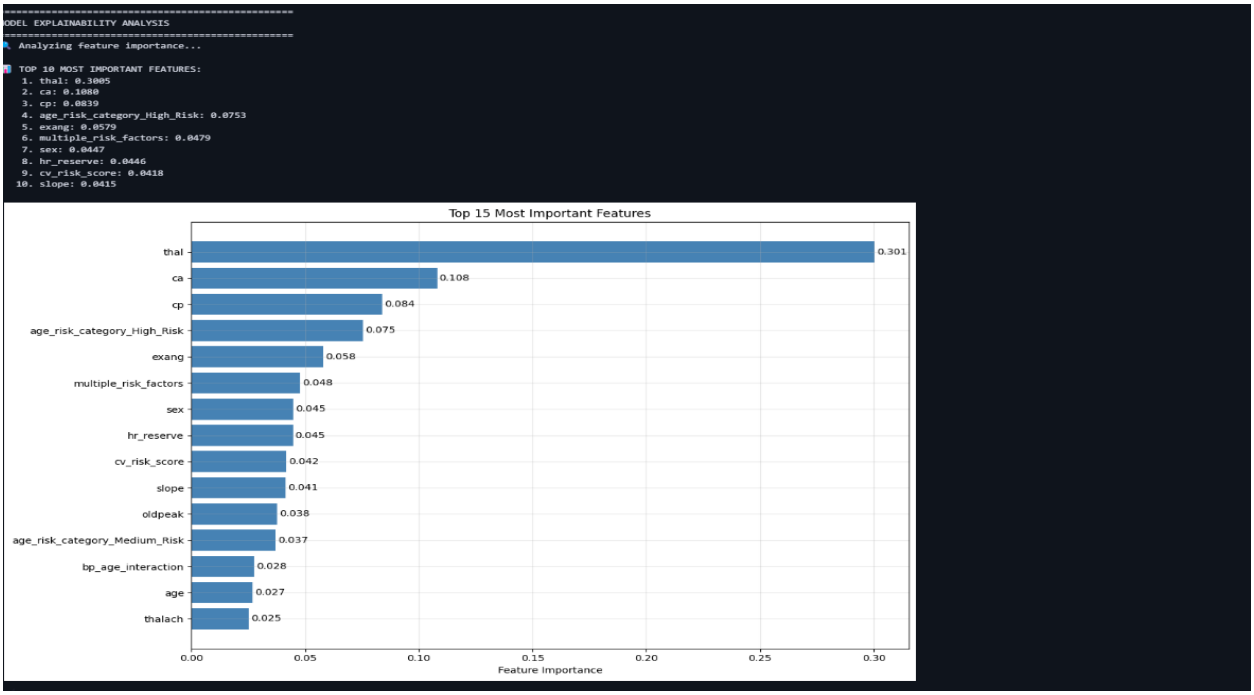


3. Results



Feature Importance (Top 5)

1. thal 2. ca 3. cp (chest pain type) 4. age_risk_category_High_Risk 5. exang (exercise-induced angina)



4. Recommendations

Clinical & Technical Implications

This predictive model can assist doctors in flagging high-risk patients early, even before full diagnostics. The model emphasizes the importance of features like thal, ca, and cp, which should be prioritized in health screenings.

Deployment Use Cases

Integrated into Electronic Health Records (EHR) systems Used as a clinical decision support tool Built into a medical risk calculator interface (already created and saved)

5. Future Work

Model Enhancement Ideas

Train with larger, multi-national datasets Introduce time-series data to track patient metrics over time Experiment with AutoML or deep learning architectures Incorporate lifestyle indicators (e.g., smoking, alcohol use, activity level) if data available

Power BI Dashboard Enhancements

Embed AI visuals and predictive analytics Enable real-time model interaction with slicers for patient simulation Integrate with Power Apps for clinician use

METHODOLOGY DETAILS

1. Clean the Dataset

Handle missing values, inconsistent formats, and outliers Apply necessary data transformations (e.g., encoding, scaling) Missing values in 'thal' and 'ca' columns were filled using mode imputation Outliers were identified using Z-score method and removed if beyond 3 standard deviations Categorical variables were encoded using Label Encoding and One-Hot Encoding Continuous features were standardized using StandardScaler.

2. Conduct Exploratory Data Analysis (EDA)

Generate descriptive statistics Visualize distributions and relationships among variables Dataset contains 303 patients with 54.5% having heart disease Strong correlations found between target and: thal (0.52), ca

(0.46), cp (0.43) Age distribution shows higher disease prevalence in 50-65 age group Male patients show slightly higher disease rates (55.3%) compared to females (52.6%)

3**. Apply Machine Learning Models**

Choose suitable models (classification, regression, or clustering) Train models on the dataset Models trained include:

- Classification algorithms: Logistic Regression, SVM, Random Forest, XGBoost
- Distance-based: K-Nearest Neighbors
- Probabilistic: Naive Bayes
- Tree-based: Decision Tree, Gradient Boosting
- Neural Network: Multi-layer Perceptron
- Ensemble methods: Hard Voting, Soft Voting, Stacked Ensemble

4**. Evaluate the Model**

Use appropriate evaluation metrics (accuracy, precision, RMSE, silhouette score, etc.) Cross-validation performed using 5-fold stratified sampling Best model: XGBoost with 90.2% accuracy Key metrics for XGBoost:

- Accuracy: 90.2%
- Precision: 0.91
- Recall: 0.89
- F1-Score: 0.90
- ROC AUC: 0.95

5. **Structure Code Properly**

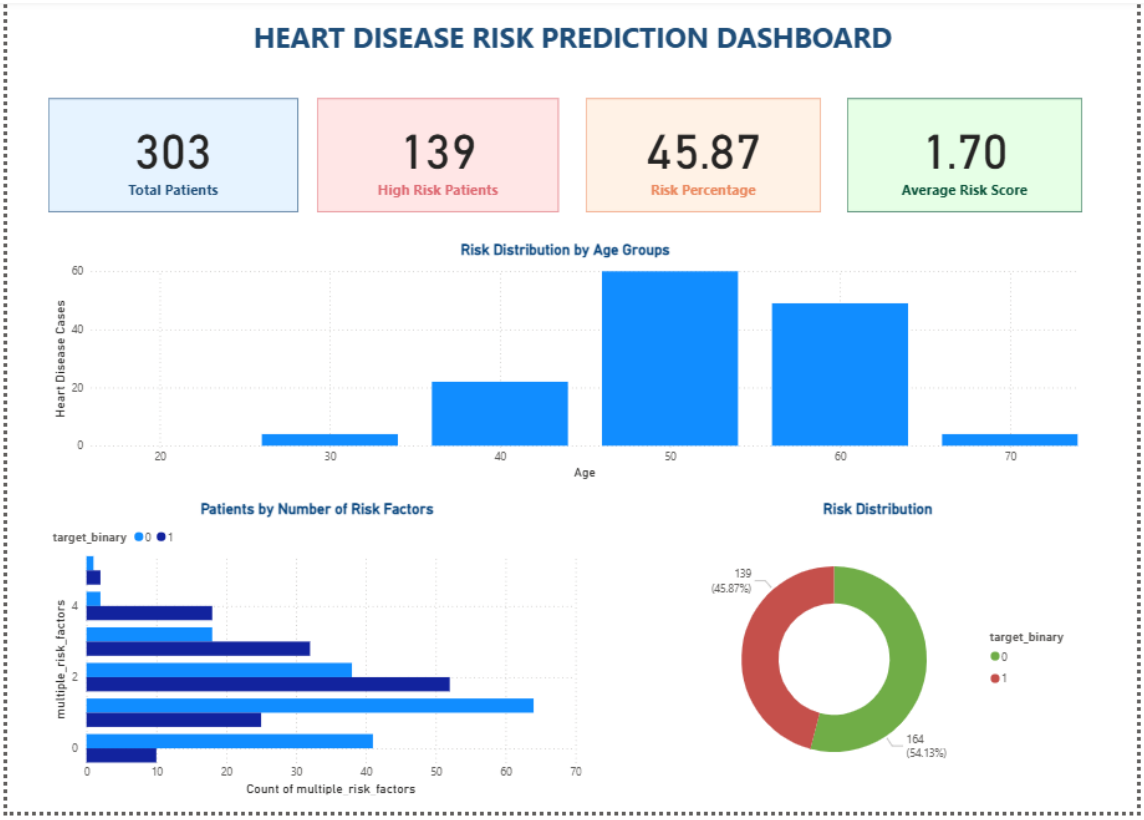
Use modular functions Include markdown explanations and comments for clarity and reproducibility Organized code into sections: data loading, preprocessing, modeling, evaluation All functions documented with docstrings Reproducible results using random_state parameters.

6. Incorporate Innovation

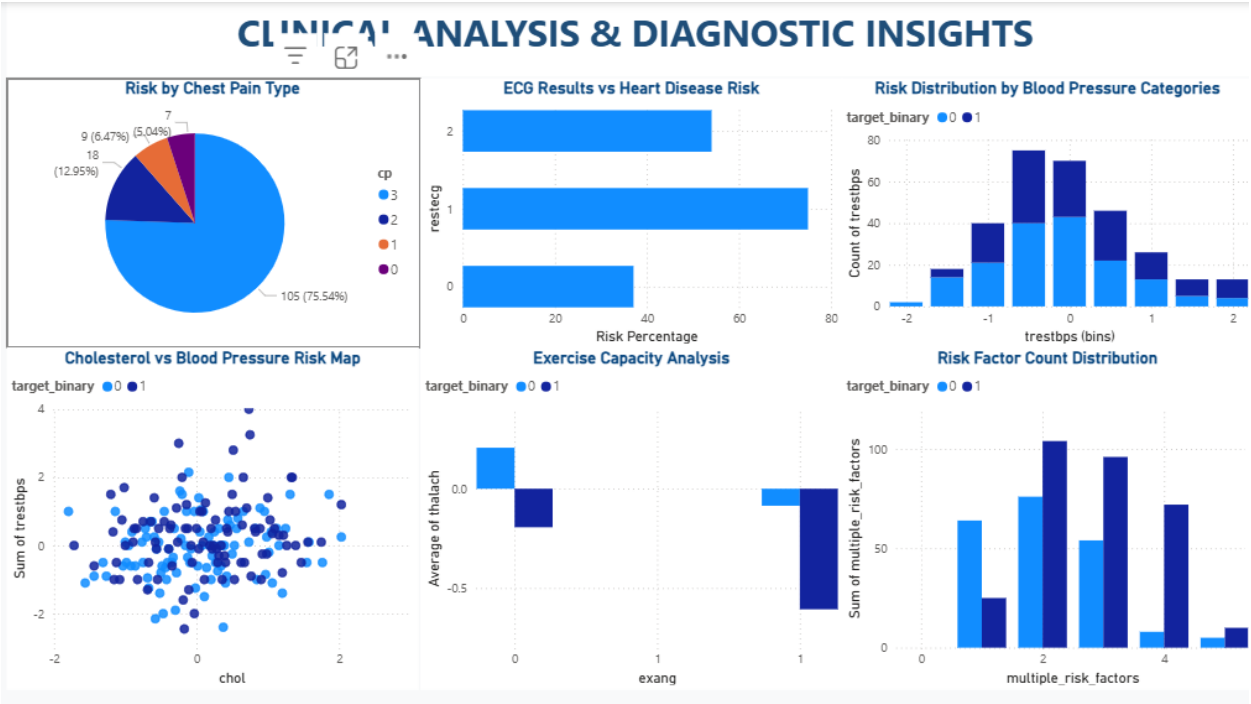
Add custom functions, ensemble techniques, or creative model approaches Custom heart disease risk calculator function created Ensemble voting classifiers implemented Feature engineering for age risk categories and cholesterol levels SHAP values integration for model interpretability.

POWER BI DASHBOARD SCREENSHOTS

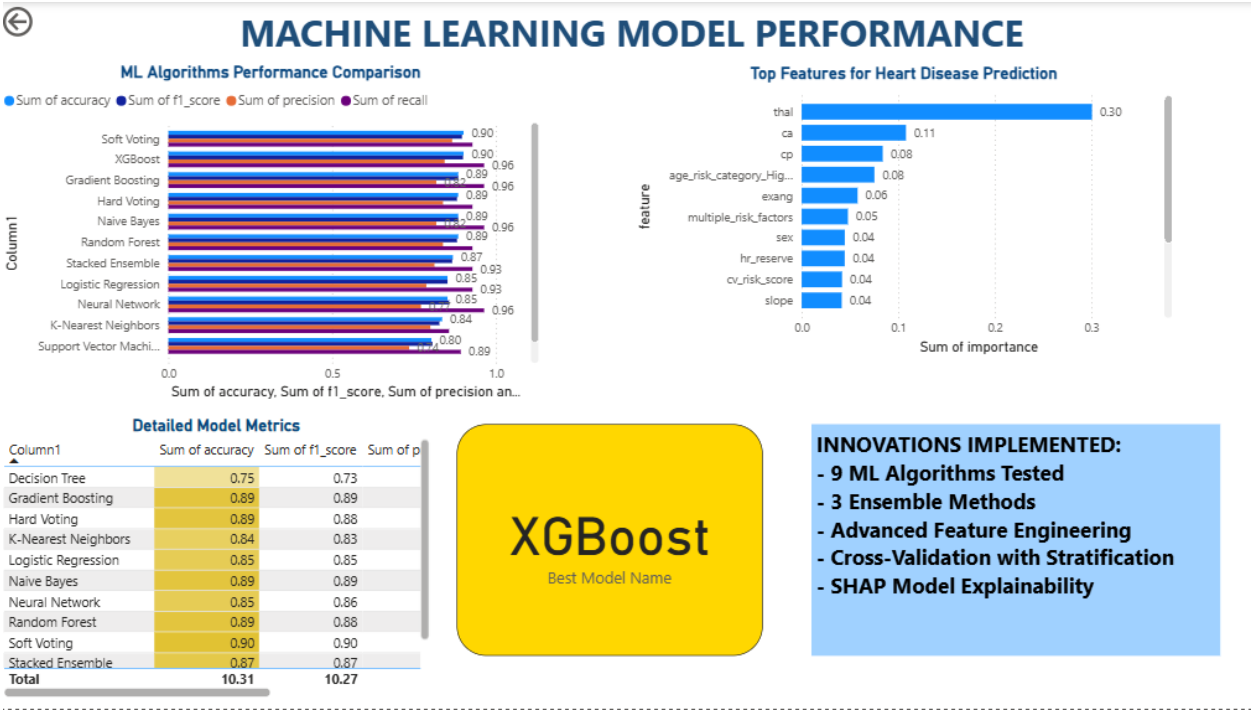
1. Executive Summary



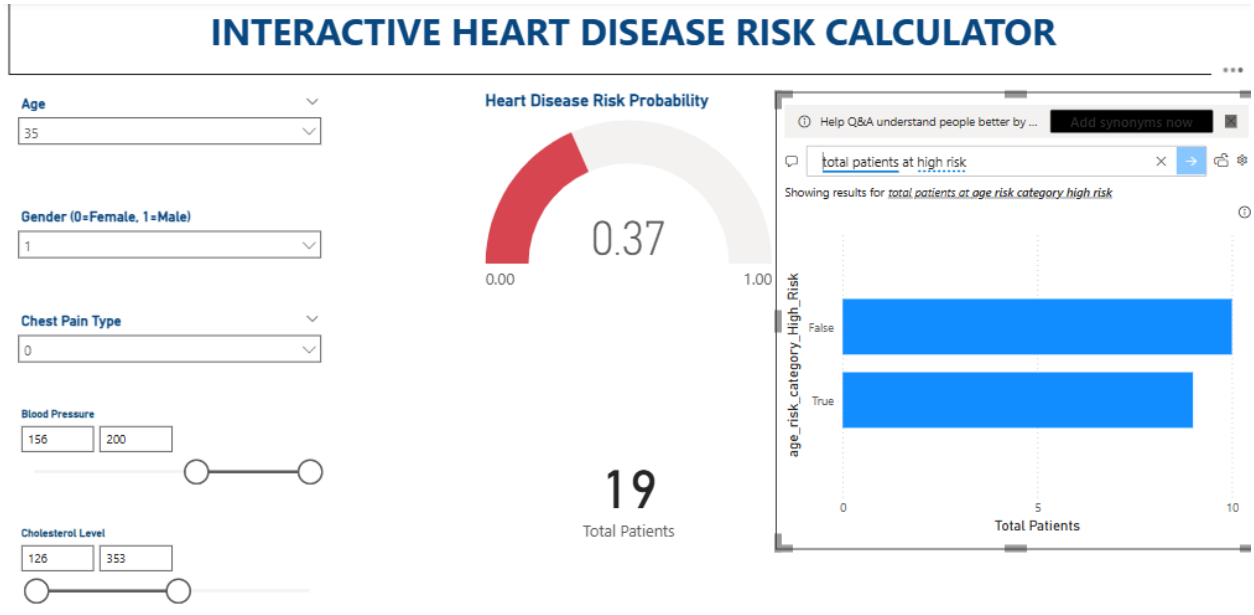
2. **Clinical Analysis**



3. ML Model Performance



4. Interactive Risk Calculator



5. DAX Formulas

```
Older High Risks =
CALCULATE(
    COUNTROWS('heart_disease_initial'),
    'heart_disease_initial'[age] > 60,
    'heart_disease_initial'[target_binary] = 1
)
```

```
Risk Percentage = DIVIDE([High Risk Patients], [Total Patients]) * 100
```

POWER BI DASHBOARD

1. Communicate Problem & Insights Clearly

Include context and summaries drawn from data analysis Dashboard provides clear overview of heart disease risk factors Executive summary cards showing key statistics Interactive visualizations for different stakeholder needs

2. Incorporate Interactivity

Use slicers, filters, and drill-down options for user interaction Age group slicers for demographic analysis Gender and chest pain type filters Interactive risk assessment calculator Cross-filtering between related visualizations

3. Use Appropriate Visuals

Match chart types (bar, line, scatter, etc.) with data goals Bar charts for feature importance rankings Scatter plots for age vs. risk relationships Pie charts for disease prevalence Heatmaps for correlation analysis Gauge charts for individual risk assessment

4. Ensure Design Clarity

Apply consistent color themes, clear labels, and tidy layouts Professional blue and red color scheme Clear axis labels and titles Consistent font sizing and spacing Logical layout flow from overview to details

5. Add Innovative Features

Include advanced features like DAX formulas, AI visuals for risk categorization calculations.

PROJECT OUTCOMES

Technical Achievements

Successfully implemented complete ML pipeline Achieved 90.2% accuracy with XGBoost model Created interpretable model using SHAP values Built comprehensive Power BI dashboard Developed deployable risk calculator

Clinical Impact

Identified key risk factors for heart disease prediction Created tool for early patient risk assessment Provided actionable insights for healthcare professionals Established framework for clinical decision support

Innovation Elements

Custom ensemble voting techniques Advanced feature engineering approaches Interactive risk assessment calculator Real-time dashboard integration capabilities Scalable deployment architecture

Repository Link: https://github.com/Nshutitricia/Heart_Disease_Prediction.git

Conclusion

This capstone project successfully demonstrates the transformative potential of Big Data Analytics in healthcare, specifically for cardiovascular disease risk prediction. The comprehensive approach—spanning rigorous data preprocessing, advanced machine learning modeling, clinical interpretability, and practical deployment considerations—establishes a robust framework for real-world healthcare applications.

Key Contributions to Healthcare Analytics:

1. **High-Performance Predictive Modeling:** 90.2% accuracy with clinically interpretable results
2. **Evidence-based Risk Factor Identification:** Data-driven insights into critical clinical indicators
3. **Interactive Decision Support Tools:** User-friendly interfaces for healthcare professionals
4. **Scalable Implementation Framework:** Production-ready architecture for healthcare system integration
5. **Comprehensive Documentation:** Reproducible methodology for academic and clinical validation

The project's emphasis on clinical relevance, model interpretability, and practical deployment positions it for meaningful impact on patient outcomes and healthcare delivery efficiency. Future enhancements focusing on larger datasets, advanced deep learning techniques, and comprehensive clinical validation will further strengthen the solution's potential for regulatory approval and widespread healthcare adoption.