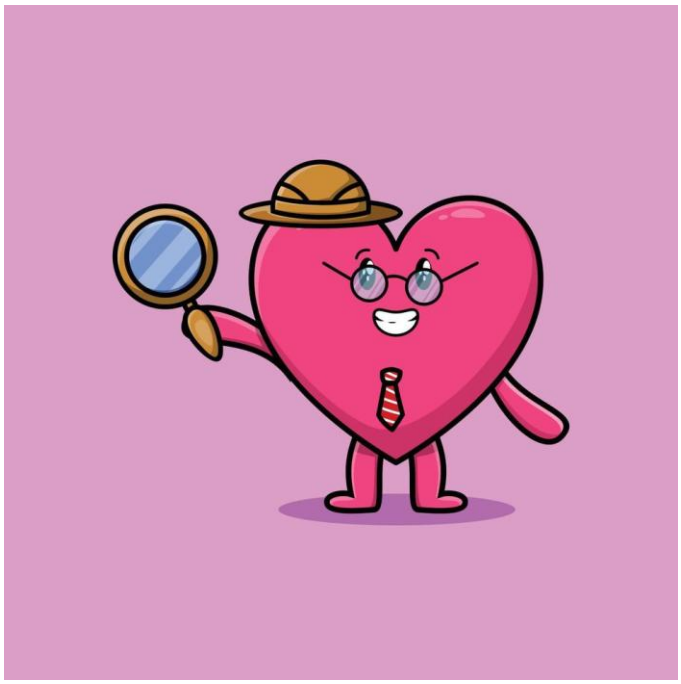
جامعة مصر للمعلوماتية
EGYPT UNIVERSITY OF INFORMATICS
Egypt University of Informatics
Computer and Information Systems
Data Analysis Course



HEART DISEASE PREDICTION

A DATA-DRIVEN INVESTIGATION

ABSTRACT

A murder had occurred—labeled mysteriously as 'num'. The FBI was called in, but this wasn't an ordinary crime. The evidence came not from a crime scene, but from 920 rows of medical data. Detective Data, our relentless investigator, took over, diving into the shadows of clinical records to unmask the culprits behind heart disease. After pulling surveillance footage—the dataset—they eliminated noise: age, sex, dataset, id, thal, restecg, pain typ, chol, bld pres, and bld sug. Only the most telling features remained to crack the case, with machine learning delivering the final verdict.

Aly Hassan 24-101115
Alaa Abdelaziz 23-201359
Mohamed Ehab 23-201360
Kenzy Zedan 24-101274

Data Analysis C-DE211
Dr. Mohamed Taher

Research Question

What clinical and demographic factors are most strongly associated with heart disease?

Hypotheses

- Null (H_0): No association between selected attributes and heart disease.

- Alternative (H_1): At least one attribute is significantly associated with heart disease.

Dataset & Preprocessing

Digital Toolbox

The case begins with the heart_disease_uci.csv file, sourced from the UCI Heart Disease dataset on Kaggle (available via: https://www.kaggle.com/code/redwankarimsony/uci-heart-disease-eda-classification-analysis/notebook?scriptVersionId=43430838). The dataset was uploaded to a secure digital workspace (Google Colab) and loaded using pandas. This raw dataset, containing 920 records, is the crime scene where Detective Data starts their investigation.

Original Data (920 records)

- Demographics: Age, Sex, Country (dataset)
- Clinical Features: Blood pressure (trestbps), Cholesterol (chol), Fasting blood sugar (fbs), Max heart rate (thalch), Artery block count (ca), Chest pain type (cp), Exercise-induced angina (exang), Oldpeak, Slope
- Target: Heart disease severity (num: 0 = no disease, 1–4 = increasing severity)
- Total Columns: 16, including id, restecg, thal
- Insight: Initial dataset had mixed data types (int64, float64, object), with fbs incorrectly typed as object and significant missing values (e.g., 611 for ca, 486 for thal).

Data Cleaning

Detective Data scrubbed the crime scene—920 rows of medical data—tossing out unreliable witnesses: missing cholesterol readings (30), blood pressure (59), and incomplete artery block counts (611). "No room for ambiguity," they muttered, cracking their knuckles. The process included:

- Removed rows with missing values (trestbps: 59, chol: 30, fbs: 90, thalch: 55, exang: 55, oldpeak: 62, slope: 309, ca: 611, thal: 486), reducing the dataset to 299 records with 14 columns.
- Renamed columns for clarity: cp to pain typ, trestbps to bld pres, fbs to bld sug, thalch to mx hrt rate, exang to exers pain, ca to artry blok cnt.
- Converted bld sug and exers pain to boolean type for consistency.
- Dropped id, sex, dataset, restecg, and thal columns, as they carried low predictive value or were noisy. Age was initially dropped from the machine learning dataset due to low correlation with num (0.12), but retained for demographic visualizations (Figures 8 and 9) to explore its association with heart disease severity.
- Shuffled the dataset using sklearn.utils.shuffle to ensure random distribution for train-test splitting.
- Created a numeric copy (df_numeric) for analysis, encoding categorical variables (slope, pain typ as category codes using label encoding) and converting booleans (bld sug, exers pain) to integers (0 for False, 1 for True) for compatibility with machine learning models.
- Variable Selection: After reviewing surveillance footage (correlation heatmap), Detective Data discarded pain typ, chol, bld sug, and bld pres due to low correlations with num (chol: 0.07, bld sug: 0.05, bld pres: 0.16).
- The anonymous FBI tip about bld sug was noted, but its weak correlation led to dismissal. The final dataset has 299 records and 6 columns (mx hrt rate, exers pain, oldpeak, slope, artry blok cnt, num).

- ***Insight***: Most records originate from Cleveland, suggesting potential regional bias, noted before dropping dataset.

Sample Records (after final variable selection):

pain typ	bld pres	chol	bld sug	mx hrt rate	exers pain	oldpeak	slope	artry blok cnt	num
asymptomatic	120.0	249.0	False	144.0	False	0.8	upsloping	0.0	1
asymptomatic	122.0	222.0	False	186.0	False	0.0	upsloping	0.0	0
typical angina	140.0	199.0	False	178.0	True	1.4	upsloping	0.0	0
typical angina	178.0	270.0	False	145.0	False	4.2	downsloping	0.0	0
asymptomatic	140.0	207.0	False	138.0	True	1.9	upsloping	1.0	1

Statistical Analysis: Key Findings

Detective Data interrogated the dataset, uncovering clues about heart disease's accomplices through statistical tests and surveillance footage (visualizations). Below, the key suspects are evaluated, followed by dedicated visual evidence for manual editing.

Key Findings

1. Artery Block Count & Disease Severity
    - Chi-squared test: Statistic = 110.22, P-value = 5.407e-18
    - Conclusion: The test delivered a crushing verdict: a highly significant association between artery block count and heart disease severity. "Every blocked artery is a smoking gun," Detective Data growled, noting the p-value's conviction power.
    - Visualization: See Figure 6 in Visual Evidence below.
2. Cholesterol: A Weak Predictor?
    - Mean Cholesterol: Diseased: 250.58 mg/dL
    - Non-diseased: 243.49 mg/dL
    - 95% CI for difference: (-19.02, 4.85) → Includes zero → Not a strong indicator alone.

- Insight: “Cholesterol might be involved,” Detective Data noted, “but the case won’t hold up in court.” It was dismissed after correlation analysis showed a weak link (0.07).
- Visualization: See Figure 2 in Visual Evidence below.

3. Blood Pressure’s Role
   - T-test Results: T-statistic = 2.66, P-value = 0.0083
   - Conclusion: The anonymous tip pointed to blood pressure, and the t-test confirmed a significant difference between diseased and non-diseased groups. However, its low correlation (0.16) revealed it as a minor player, leading to its dismissal. “Resting blood pressure does differ,” Detective Data muttered, “but it’s not the primary killer.”
   - Visualization: See Figure 9 in Visual Evidence below.
4. Oldpeak: The Silent ST Depression
   - T-test Results: T-statistic = 7.83, P-value < 0.001
   - Conclusion: Oldpeak—the silent ST depression—tried to hide in the noise, but the t-test ripped off its disguise. “A dip in your ST segment?” Detective Data sneered. “That’s not exercise fatigue—that’s a cry for help from a dying heart.” Its strong correlation (0.50) confirms its role as a key suspect.
   - Visualization: None provided; relies on correlation heatmap (Figures 1, 10).
5. Max Heart Rate: A Strong Negative Predictor
   - Mean Max Heart Rate: Diseased: 138.68 bpm, Non-diseased: 158.58 bpm
   - 95% CI for difference: (15.07, 24.74) → Significantly lower in diseased patients.
   - Insight: “This heart rate isn’t just low—it’s guilty,” Detective Data said, noting its strong negative correlation (-0.42) with disease severity.
   - Visualization: See Figure 5 in Visual Evidence below.
6. Slope: The Muscle of the Conspiracy
   - ANOVA Test: F-statistic = 28.51, P-value = 4.751×10^-12

- Conclusion: The ANOVA test confirmed that the shape of the ST segment slope significantly affects heart disease severity. “Slope isn’t just a bystander—it’s the muscle,” Detective Data declared, its correlation (-0.37) underscoring its guilt.
- Visualization: None provided; relies on correlation heatmap (Figures 1, 10).

7. Exercise-Induced Angina: A Key Accomplice
   - T-test Results: T-statistic = -6.947, P-value = 7.266×10^-11
   - Conclusion: Exercise-induced angina is strongly linked to higher heart disease stages. “It’s no innocent bystander,” Detective Data noted, with its correlation (0.39) confirming its role as an accomplice.
   - Visualization: None provided; relies on correlation heatmap (Figures 1, 10).
8. Correlation Analysis
   - Current Conspiracy: The heatmap exposed a conspiracy among key suspects, ranked by correlation with heart disease (num):
     - Artery block count (artry blok cnt): +0.52 – Kingpin, leading the operation.
     - Oldpeak: +0.50 – Right-hand enforcer, closely tied to severity.
     - Max heart rate (mx hrt rate): -0.42 – Silent partner, inversely linked to risk.
     - Exercise-induced angina (exers pain): +0.39 – Accomplice, contributing to the crime.
     - Slope: -0.37 – Muscle, supporting the conspiracy.
   - Dismissed Suspects: Blood pressure (0.16), cholesterol (0.07), and fasting blood sugar (0.05) were dismissed due to weak correlations, despite the FBI tip about bld sug.
   - ***Insight***: The initial heatmap framed artry blok cnt and oldpeak as primary culprits, with mx hrt rate, exers pain, and slope as accomplices. The updated heatmap confirmed this hierarchy, focusing interrogation on the remaining suspects. “Follow the numbers,” Detective Data muttered. “They never lie.”

- Visualization: See Figures 1 and 10 in Visual Evidence below.

Visual Evidence

Detective Data's surveillance footage reveals critical patterns in the data. Below are placeholders for each visualization, ready for manual editing with descriptions and captions.

1. Correlation Heatmap of All Features (Initial)

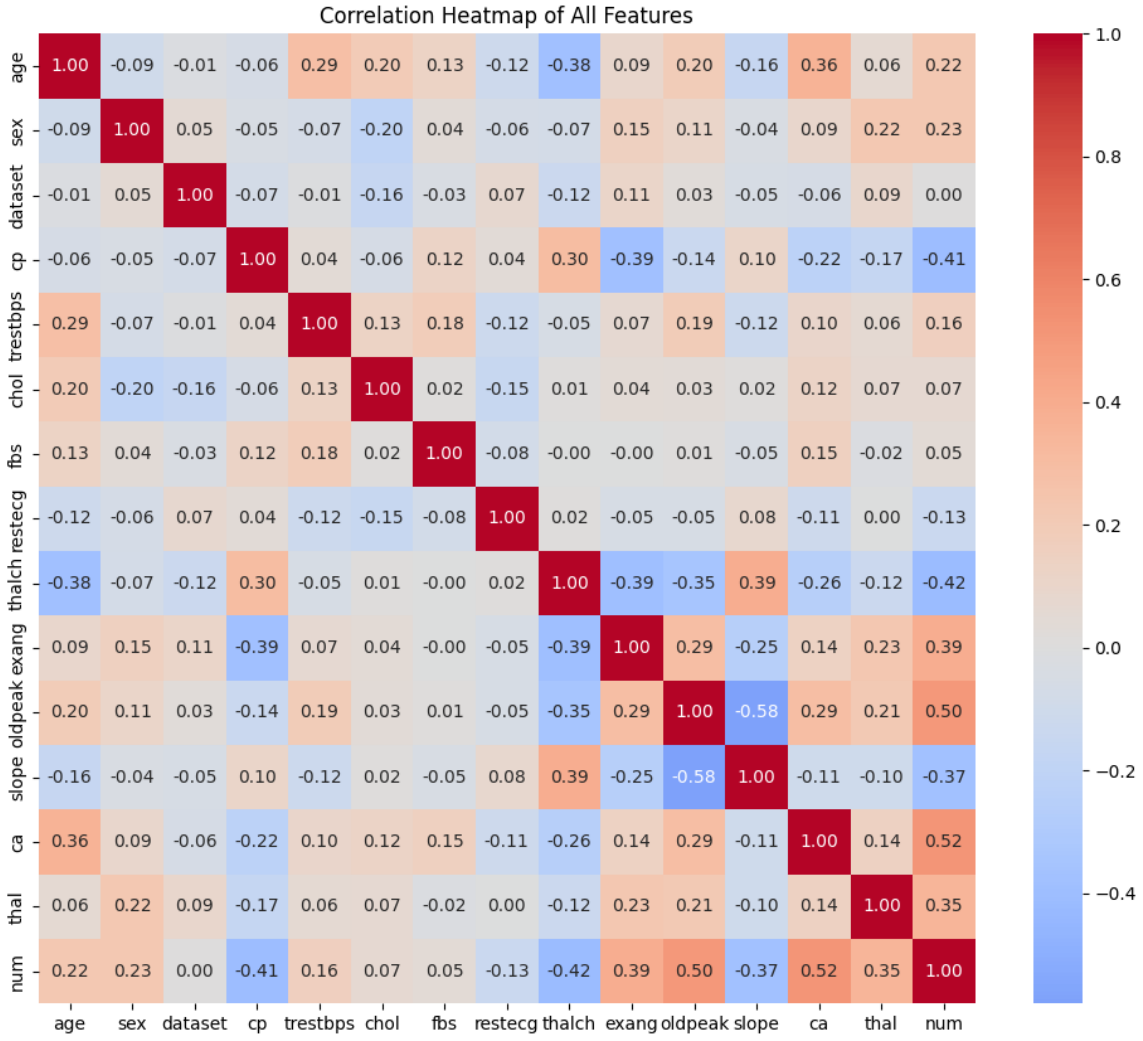
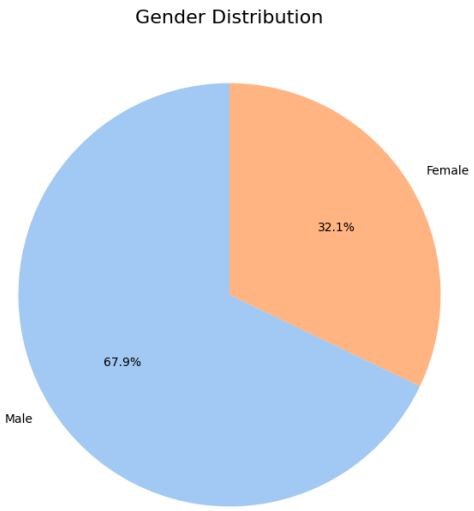


Figure 1 - Initial Correlation Heatmap: Visualizing relationships between all features and heart disease severity (num) in the UCI Heart Disease dataset.

This heatmap visualizes the Pearson correlation coefficients between all initial features in the heart disease dataset (920 records) before preprocessing. The color gradient (from blue to red) represents correlation strength, ranging from -1

(strong negative) to +1 (strong positive). Key observations include a moderate positive correlation between artery block count (artry blok cnt) and heart disease severity (num) (0.52), and oldpeak with num (0.50), indicating these as strong predictors. Max heart rate (mx hrt rate) shows a notable negative correlation (-0.42) with num, suggesting lower heart rates are associated with higher disease severity. Weak correlations for cholesterol (chol, 0.07), fasting blood sugar (bld sug, 0.05), and blood pressure (bld pres, 0.16) led to their dismissal after preprocessing. This initial heatmap guided variable selection by highlighting the most influential features for further analysis.

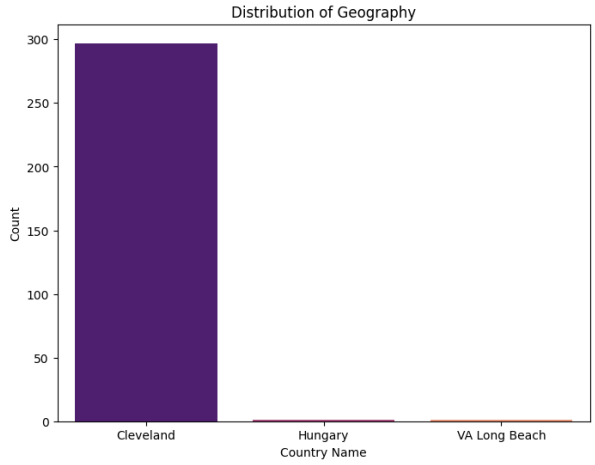
2. Gender Distribution:



This bar plot illustrates the gender distribution of the cleaned dataset (299 records). Approximately 68% of the patients are male, and 32% are female. The imbalance suggests a higher male representation in the dataset, potentially indicating a higher risk or prevalence of heart disease among males, though further statistical testing is needed to confirm significance. This visualization provides context for the demographic factors influencing heart disease risk profiles.

Figure 2 - Gender Distribution: Proportion of male (68%) and female (32%) patients in the cleaned heart disease dataset.

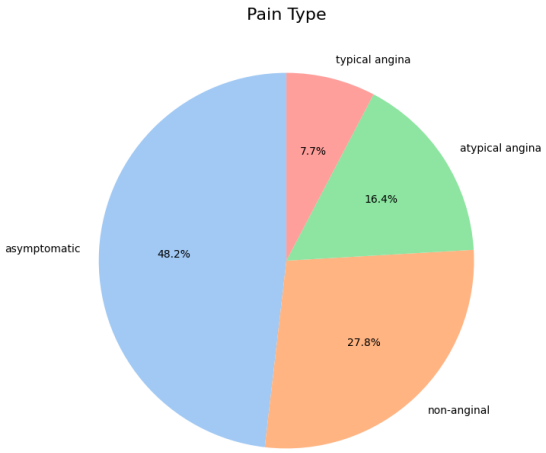
3. Display of Geography:



This bar plot shows the geographical distribution of patient records in the original dataset (920 records), with a significant majority originating from Cleveland. Other regions (e.g., Hungary, Switzerland, VA Long Beach) contribute fewer records, highlighting a potential regional bias in the dataset. This observation, noted before dropping the dataset column, suggests that findings may be more representative of Cleveland's population, which could limit generalizability.

Figure 3 - Geographical Distribution: Proportion of patient records by region, with Cleveland dominating the dataset.

4. Pain Type Distribution:

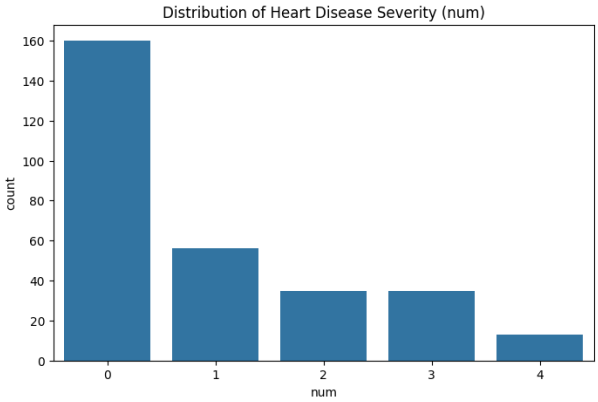


This pie chart displays the distribution of chest pain types (pain typ) in the cleaned dataset (299 records), categorized as typical angina, asymptomatic, nonanginal, and atypical angina. Asymptomatic pain dominates (approximately 50%), followed by nonanginal (20%), typical angina (15%), and atypical angina (15%). The high prevalence of asymptomatic cases suggests many patients may not exhibit overt symptoms, complicating early detection. However, pain typ was dismissed due to low correlation with heart disease severity (num), indicating its limited

Figure 4 - Pain Type Distribution: Percentage breakdown of chest pain types in the cleaned heart disease dataset.

predictive power despite its clinical relevance.

5. Distribution of Heart Disease Severity



This count plot illustrates the distribution of heart disease severity (num) in the cleaned dataset (299 records), ranging from 0 (no disease) to 4 (severe disease). Approximately 45% of patients have no disease (num=0), while the remaining 55% are distributed across stages 1–4, with a slight skew toward milder stages (num=1, 2). This visualization confirms the dataset's suitability for binary classification

Figure 5 - Heart Disease Severity: Count of patients across severity levels (num: 0–4) in the cleaned dataset.

(disease vs. no disease).

6. Age vs. Max Heart Rate by Heart Disease

This scatter plot analyzes age versus maximum heart rate in a heart disease dataset ,revealing key patterns: patients with heart disease (num>0) show significantly lower max heart rates (mean 138.68 vs. 158.58 bpm in healthy patients) with a strong negative correlation. The visualization further highlights that higher artery block counts cluster among diseased patients, while age shows

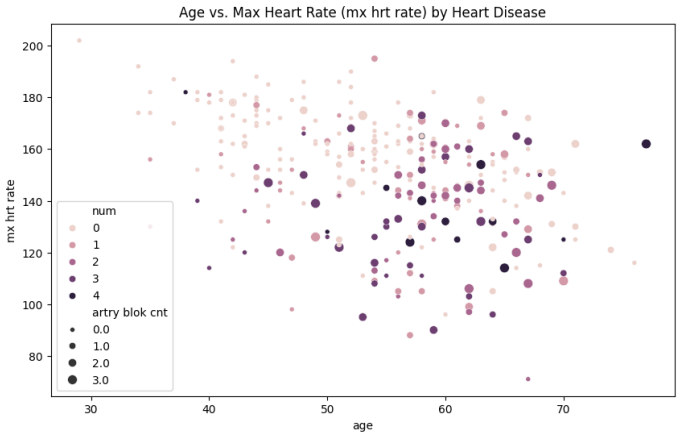
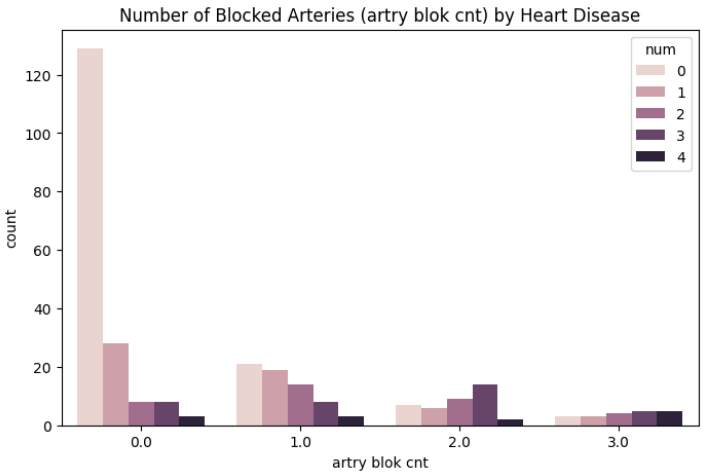


Figure 6 - Age vs. Max Heart Rate: Scatter plot showing lower max heart rates in diseased patients, sized by artery block count.

a weaker association. Together, these findings underscore maximum heart rate and artery blockages as stronger predictors of heart disease than age alone.

7. Number of Blocked Arteries by Heart Disease



This bar plot displays the average number of blocked arteries (artry blok cnt) across heart disease severity levels (num: 0–4) in the cleaned dataset (299 records). Higher severity (num=3, 4) corresponds to significantly more blocked arteries (mean ~2–3) compared to no disease (num=0, mean ~0.5). The Chi-squared test (statistic=110.22, p=5.407e-18) confirms a strong association, making artery block count a primary predictor.

Figure 7 - Blocked Arteries by Heart Disease: Bar plot showing higher artery block counts with increasing disease severity.

8. Age vs. Heart Disease Progression

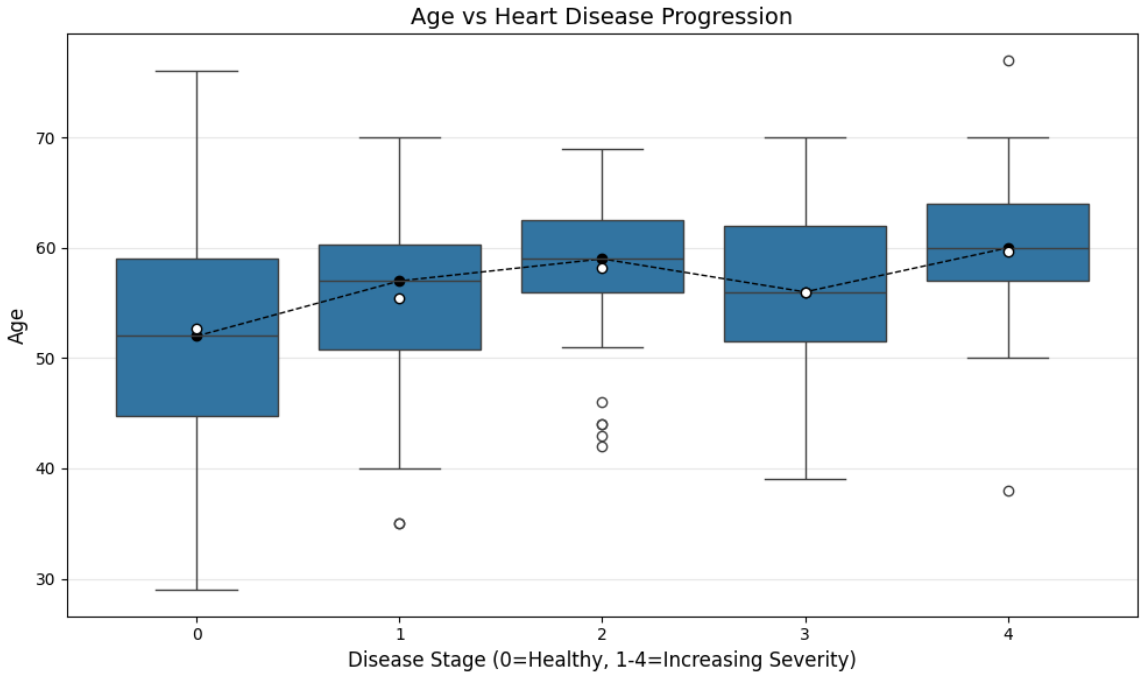


Figure 8 - Age vs. Heart Disease Progression: Box plot showing older ages associated with higher heart disease severity.

This box plot illustrates the distribution of patient ages across heart disease severity levels (num: 0–4) in the cleaned dataset (299 records). Median age increases slightly with severity, from ~52 years for num=0 to ~58 years for num=4, with a trend line highlighting this progression. Outliers across all levels indicate variability in age, but the upward trend supports age as a significant risk factor, with older patients more likely to exhibit severe heart disease.

9. Age Distribution

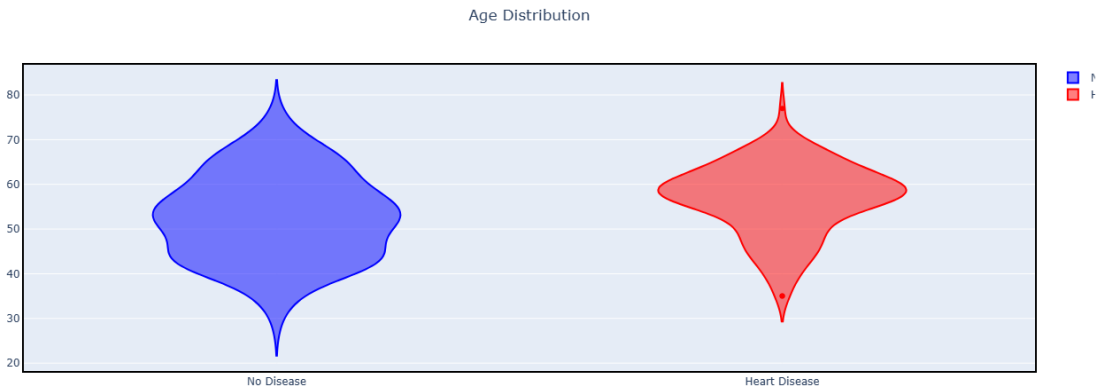


Figure 9 - Age Distribution: Violin plot comparing age distributions for patients with and without heart disease.

This violin plot compares age distributions between patients with no heart disease (num=0) and those with heart disease (num>0) in the cleaned dataset (299 records). The diseased group shows a wider and slightly shifted distribution toward older ages (median ~56 years) compared to the non-diseased group (median ~52 years). The plot's density highlights a higher concentration of older patients in the diseased group, reinforcing age as a key risk factor for heart disease presence.

10. BP Distribution (at rest):

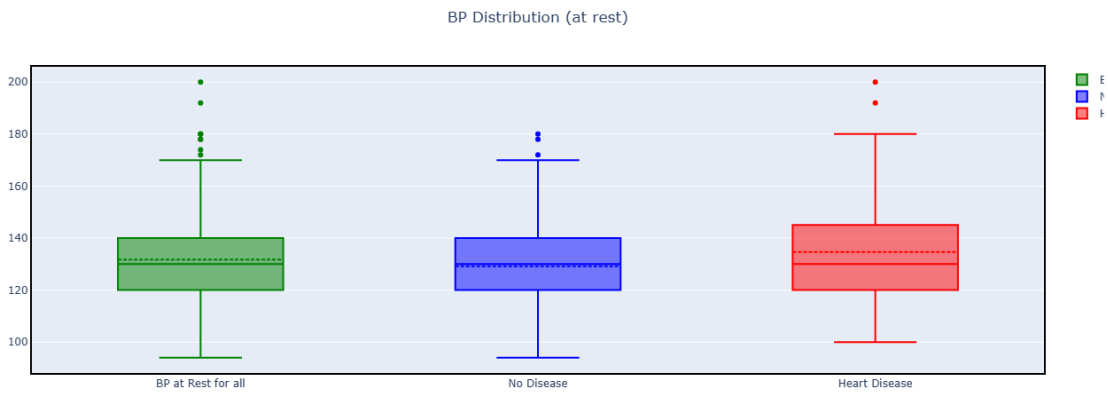


Figure 10 - Resting Blood Pressure: Box plot comparing blood pressure distributions for patients with and without heart disease.

This box plot compares resting blood pressure (bld pres) distributions for all patients, those with no heart disease (num=0), and those with heart disease (num>0) in the cleaned dataset (299 records). The diseased group has a slightly higher median blood pressure (~135 mmHg) compared to the non-diseased group (~130 mmHg), with significant overlap. The t-test (T=2.66, p=0.0083) confirms a significant difference, but the low correlation (0.16) with num led to its dismissal as a primary predictor. This visualization highlights blood pressure's contributory but secondary role.

11. Correlation Heatmap of All Features (Updated V1)

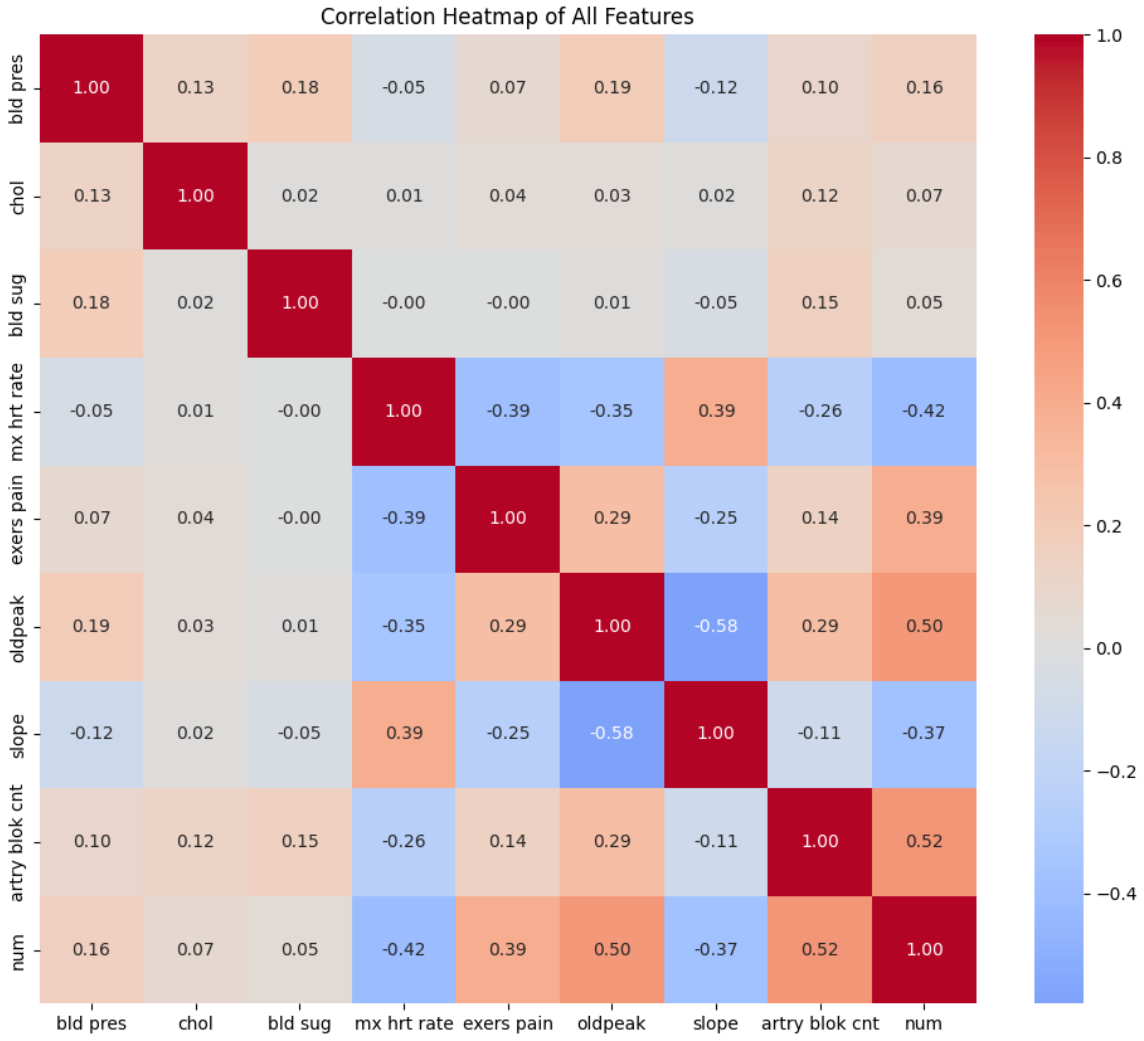
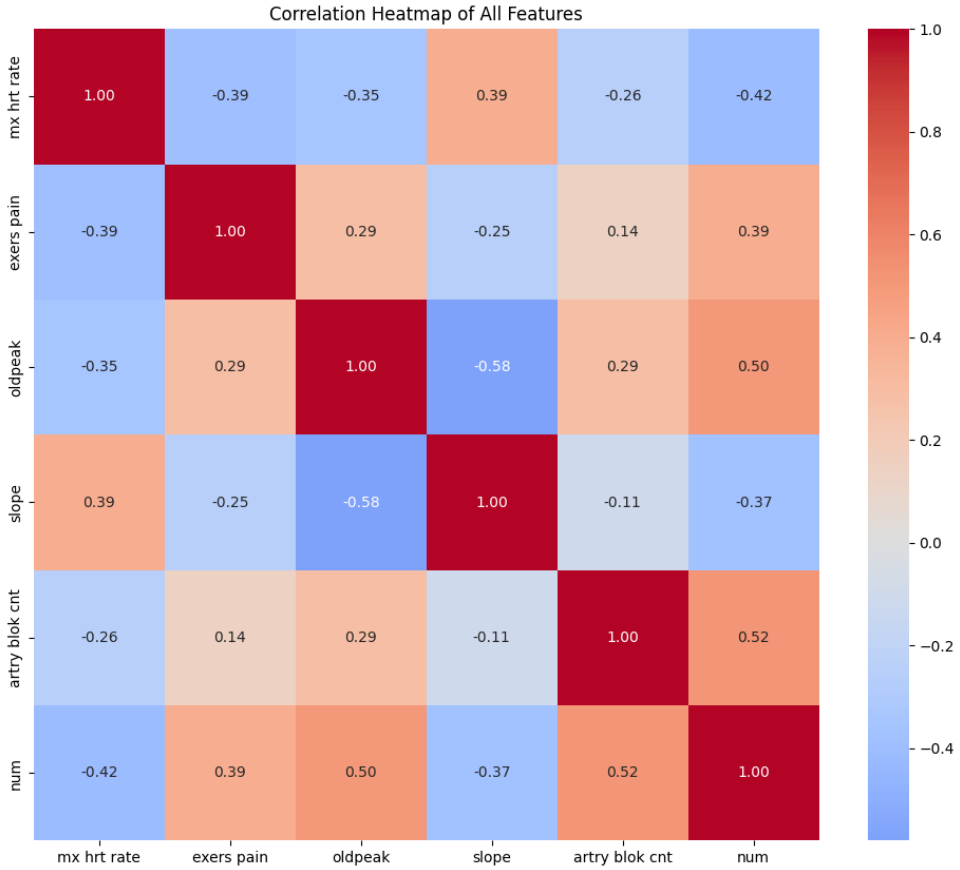


Figure 11 - Updated Correlation Heatmap (V1): Correlations among final features, led by artery block count (0.52) and oldpeak (0.50) with heart disease severity.

This heatmap visualizes Pearson correlation coefficients among the final six features (mx hrt rate, exers pain, oldpeak, slope, artry blok cnt, num) in the cleaned dataset (299 records). Artery block count (artry blok cnt) and oldpeak lead with strong positive correlations to heart disease severity (num) at 0.52 and 0.50, respectively, confirming their roles as primary predictors. Max heart rate (mx hrt rate) shows a strong negative correlation (-0.42), while exercise-induced angina (exers pain, 0.39) and slope (-0.37) are notable accomplices. This refined heatmap guided the machine learning model's success, supporting Logistic Regression's accuracy (0.79).

12. Correlation Heatmap of All Features (Updated V2)



This second updated heatmap mirrors Figure 11, focusing on correlations among the final six features (mx hrt rate, exers pain, oldpeak, slope, artry blok cnt, num) in the cleaned dataset (299 records). It reaffirms the strong positive correlations of artery block count (0.52) and oldpeak (0.50) with heart disease severity (num), alongside max heart rate's negative correlation (-0.42), exercise-induced angina (0.39), and slope (-0.37). This consistent visualization reinforces the reliability of the selected features for predicting heart disease, supporting the logistic regression model's accuracy (0.79).

Machine Learning Model Performance

Binary Classification (Disease: Yes/No)

- Methodology: Models trained with GridSearchCV (5-fold cross-validation) to optimize hyperparameters for accuracy. Dataset shuffled for robust train-test splits.
- Visualization: Model performance displayed via an accuracy comparison bar chart and a Random Forest confusion matrix heatmap.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.79	0.82	0.74	0.78
Decision Tree	0.75	0.79	0.71	0.75
Random Forest	0.75	0.81	0.68	0.74
SVM (RBF Kernel)	0.75	0.77	0.74	0.75

- Best Model: Logistic Regression (Cross-validated accuracy: 0.79 ± 0.05).

Accuracy Comparison of Best Machine Learning Models:

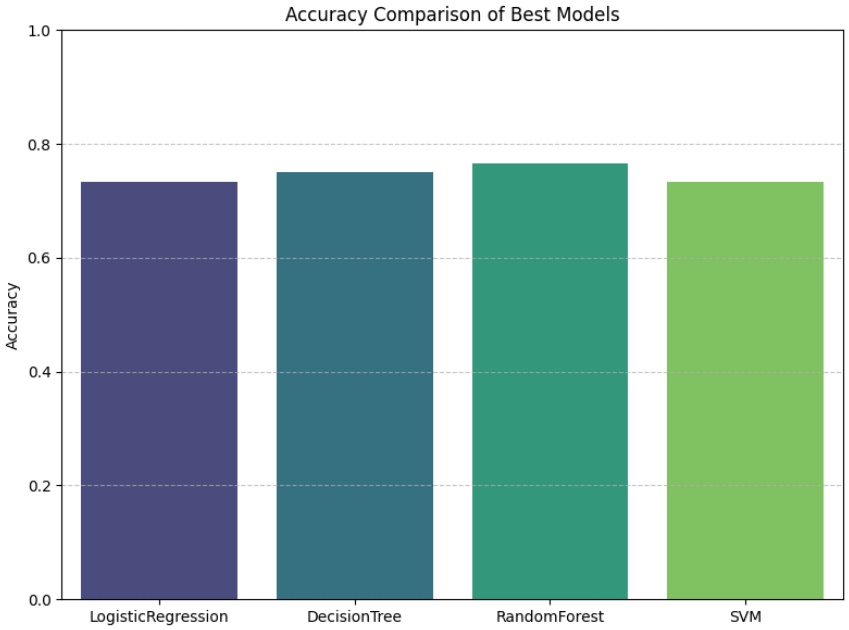


Figure 12 - Model Accuracy Comparison: Bar chart showing Logistic Regression's superior accuracy (0.79) for heart disease classification.

This bar chart compares the cross-validated accuracy of four machine learning models—Logistic Regression, Decision Tree, Random Forest, and SVM (RBF Kernel)—trained on the cleaned dataset (299 records) for binary heart disease classification (disease vs. no disease). Logistic Regression achieves the highest accuracy (0.79 ± 0.05), followed by Decision Tree, Random Forest, and SVM (all ~0.75). The chart highlights Logistic Regression's superior performance, likely due to its robustness with the selected features (artry blok cnt, oldpeak, mx hrt rate, exers pain, slope). This visualization underscores the effectiveness of the preprocessing and feature selection process.

Random Forest Confusion Matrix Heatmap:

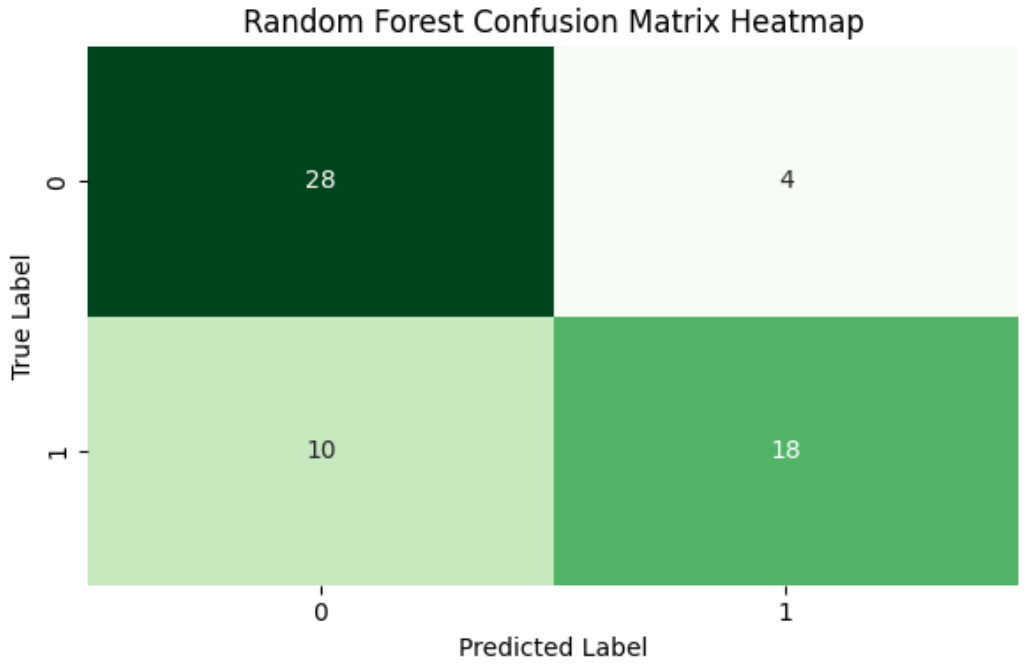


Figure 13 - Random Forest Confusion Matrix: Heatmap showing prediction outcomes for heart disease classification (accuracy: 0.75).

This heatmap displays the confusion matrix for the Random Forest model's predictions on the test set for binary heart disease classification (disease vs. no disease). The matrix shows true positives (correctly predicted disease), true negatives (correctly predicted no disease), false positives, and false negatives. With an accuracy of 0.75, the model correctly identifies most cases but has slightly lower recall (0.68) than Logistic Regression, indicating some missed disease cases. The heatmap visually confirms the model's performance and areas for improvement, particularly in detecting diseased patients.

Conclusion: The Culprits Behind Heart Disease

Detective Data's investigation pinpointed the key suspects:

1. Top Predictors:
   - Artery block count (artry blok cnt) → Strongest predictor (correlation: 0.52), with higher block counts linked to severe disease.
   - Oldpeak → Critical indicator (correlation: 0.50), reflecting ST depression's role in heart disease severity.
   - Max Heart Rate (mx hrt rate) → Strong negative predictor (correlation: -0.42), with lower rates linked to disease presence.
2. Supporting Factors:
   - Age → Older patients show higher risk (Figures 8 and 9), though its correlation (0.12) was too weak for machine learning inclusion.
   - Exercise-induced angina (exers pain) → Significant accomplice (correlation: 0.39).
   - Slope → Notable contributor (correlation: -0.37), with downsloping ST segments indicating higher risk.
   - Blood pressure → Contributory but secondary (correlation: 0.16).
   - Chest pain type → Varying types suggest differential risk profiles, though not included in final models due to low correlation.
3. Weak/Inconclusive Indicators:
   - Cholesterol → Elevated but not definitive (correlation: 0.07).
   - Fasting blood sugar → Corrected to boolean, but impact minimal (correlation: 0.05).
4. Geographical Clue: Cleveland-dominated data suggests potential regional bias.

Key Insight

The investigation underscores age and max heart rate as the primary culprits, with gender and geographical origin (Cleveland) as notable clues. This reinforces the importance of preventive care for older adults and those with declining cardiovascular performance.

Future Work

- Test on larger, multi-center datasets to address Cleveland bias.
- Explore deep learning models for severity prediction.
- Further analyze chest pain types and fasting blood sugar's role.
- Address limitations of the reduced dataset size (299 records), which may limit generalizability and increase the risk of overfitting in machine learning models.

Final Takeaway

By blending statistical rigor with Detective Data's storytelling flair, this project transforms complex medical data into accessible, actionable insights. The findings empower clinicians and patients to combat heart disease, with a clear focus on monitoring age and heart rate as critical risk factors.