

Predictive Analytics for Cardiovascular Disease Detection: A Machine Learning Approach Using Spark MLlib

- Avanti Chandratre
Chhaya Tundwal

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for approximately 31% of global mortality. Early detection and proactive management are critical to reducing CVD-related fatalities, particularly for high-risk individuals. Traditional diagnostic methods are often time-consuming and require significant medical expertise. By leveraging big data processing and predictive analytics, healthcare providers can develop a machine-learning-based model to assess a patient's risk of heart disease in real time.

1.1. Research Questions

- How can machine learning and big data analytics improve the early detection of cardiovascular diseases by accurately predicting individuals at high risk based on key health indicators?
- How does **age influence** the likelihood of developing heart disease across different gender groups?
- Does the **presence of chest pain and exercise-induced angina** increase the risk of heart disease, and how do these factors interact with other variables?
- What role do **resting electrocardiogram (ECG)** results play in identifying individuals at high risk for heart disease?
- What patterns exist between old peak depression (from exercise) and heart disease occurrence, and do these patterns differ based on other clinical features?

1.2. Data Selection and Definition

The dataset belongs to the Healthcare domain. It helps us predict heart failure based on the patients other measurable medical parameters. The dataset consists of 11 key features that can be used to predict the likelihood of cardiovascular disease (CVD). Each attribute provides critical health indicators that contribute to risk assessment. Below is a breakdown of the attributes and their significance:

- a) **Age (years)**: Represents the patient's age, a major risk factor for heart disease.
- b) **Sex (M/F)**: Identifies the gender of the patient, as men and women have different risk levels for CVD.
 - **ChestPainType (TA, ATA, NAP, ASY)**: Categorizes the type of chest pain experienced by the patient, which can indicate different levels of heart disease risk:
 - **TA (Typical Angina)**: Chest pain due to reduced blood flow to the heart.
 - **ATA (Atypical Angina)**: Chest pain with non-classical symptoms.
 - **NAP (Non-Anginal Pain)**: Chest pain not related to the heart.
 - **ASY (Asymptomatic)**: No chest pain, but may still have heart disease.

c) **RestingBP** (*mm Hg*): Measures the resting blood pressure. High values indicate hypertension, a major risk factor for CVD.

d) **Cholesterol** (*mg/dl*): Represents serum cholesterol levels. High cholesterol contributes to plaque buildup in arteries, leading to heart disease.

e) **FastingBS** (*0/1*): Binary indicator of fasting blood sugar levels:

- **1:** Fasting blood sugar > 120 mg/dl (indicates diabetes or prediabetes, which increases heart disease risk).
- **0:** Normal fasting blood sugar.

f) **RestingECG** (*Normal, ST, LVH*): Results from a resting electrocardiogram, which detects heart abnormalities:

- **Normal:** No significant issues detected.
- **ST:** ST-T wave abnormalities, which may indicate ischemia or infarction.
- **LVH:** Left ventricular hypertrophy, suggesting heart strain or chronic high blood pressure.

g) **MaxHR** (*Numeric, 60–202*): Maximum heart rate achieved during exercise. Lower values may indicate poor heart function.

h) **ExerciseAngina** (*Y/N*): Identifies whether exercise-induced angina (chest pain) occurs:

- **Y (Yes):** Indicates potential heart disease.
- **N (No):** No angina detected during exercise.

i) **Oldpeak** (*Numeric*): Measures ST segment depression in an ECG, which helps assess ischemia (reduced blood flow to the heart). Higher values may indicate more severe heart disease.

j) **ST_Slope** (*Up, Flat, Down*): Describes the slope of the ST segment during peak exercise, which helps assess heart function:

- **Up:** Normal response.
- **Flat:** May indicate ischemia.
- **Down:** Suggests more severe heart disease.

k) **HeartDisease** (*0/1*): The target variable indicating heart disease presence:

- **1:** Indicates heart disease.
- **0:** Normal (no heart disease detected).

This dataset allows for predictive modeling using machine learning to classify individuals based on their risk of cardiovascular disease, aiding in early detection and intervention strategies.

2. Data Ingestion & Preprocessing using Spark

- Number of rows in the dataset: 8
- Number of unique values for each parameter:

age	sex	chest_pain_type	resting_bp	cholesterol	fasting_bs	resting_ecg	max_hr	exercise_angina	oldpeak	st_slope	heart_disease
-----	-----	-----------------	------------	-------------	------------	-------------	--------	-----------------	---------	----------	---------------

Unique Values	50	2	4	67	222	2	3	119	2	53	3	2
Missing Values	0	0	0	0	0	0	0	0	0	0	0	0

- There are no missing values in the dataset and we dropped the duplicate rows if present
- Applied categorical encoding techniques like one-hot encoding

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution and relationships of key variables in the dataset. Key analyses include:

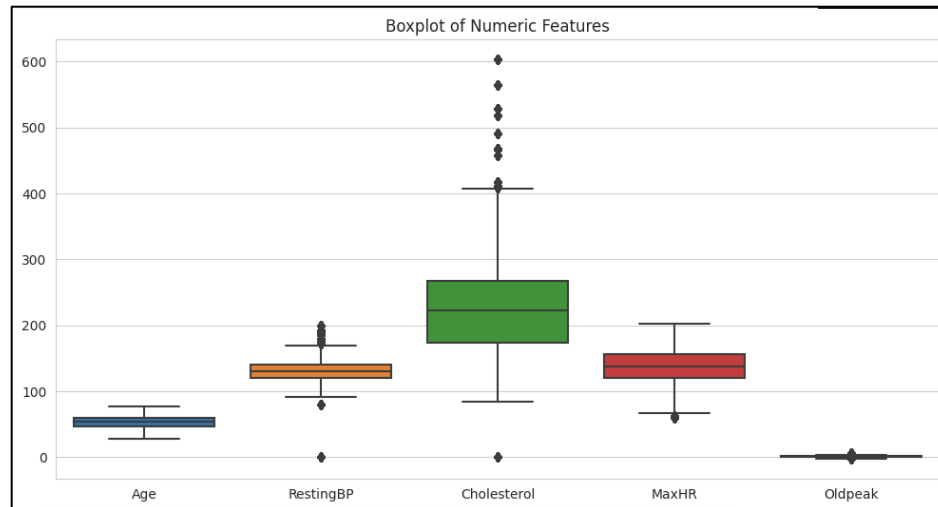
3.1. Summary statistics using Spark queries.

Feature	Minimum	Maximum	Average	Standard Deviation
Age	28	77	53.51	9.432
bp	0	200	132.4	18.51
chol	0	603	198.8	109.4
hr	60	202	136.81	25.5
oldpeak	-2.6	6.2	0.88	1.1

- **Age Metrics:**
 - The patients range from 28 to 77 years old
 - Average age is about 53.5 years
 - Standard deviation of 9.43 years indicates moderate variation in age
- **Blood Pressure (BP) Metrics:**
 - Range from 0 to 200 mmHg (the 0 value is likely an error in data collection)
 - Average BP is 132.4 mmHg, which is in the elevated/high blood pressure range
- **Cholesterol Metrics:**
 - Range from 0 to 603 mg/dL (the 0 value is likely an error in data recording)
 - Average cholesterol level is 198.8 mg/dL, which is borderline high
 - Very high standard deviation (109.4) indicates extreme variation in cholesterol levels
- **Heart Rate (HR) Metrics:**
 - Range from 60 to 202 BPM
 - Average HR is 136.81 BPM, which is elevated (normal resting HR is typically 60-100 BPM)

3.2. Visualizations and Insights

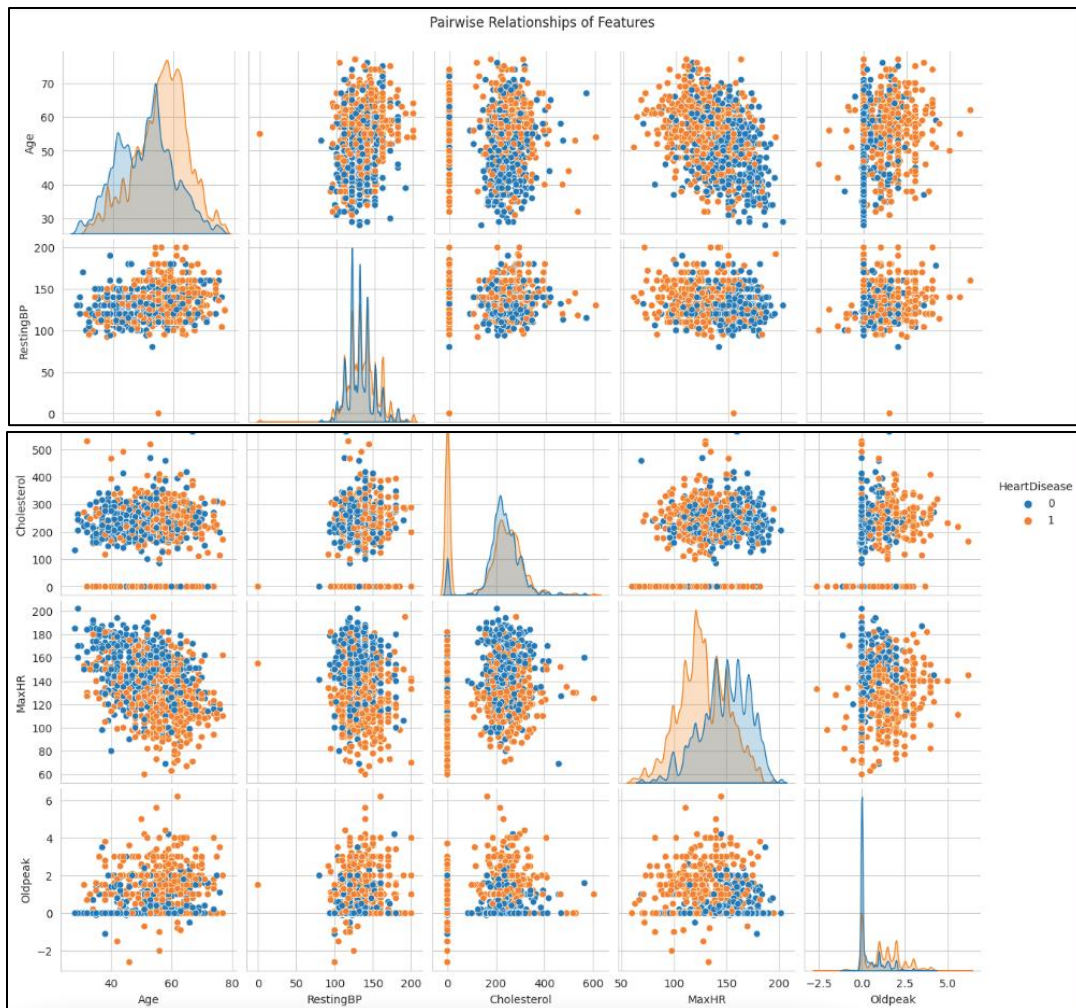
A. Comparing numeric features to understand outliers



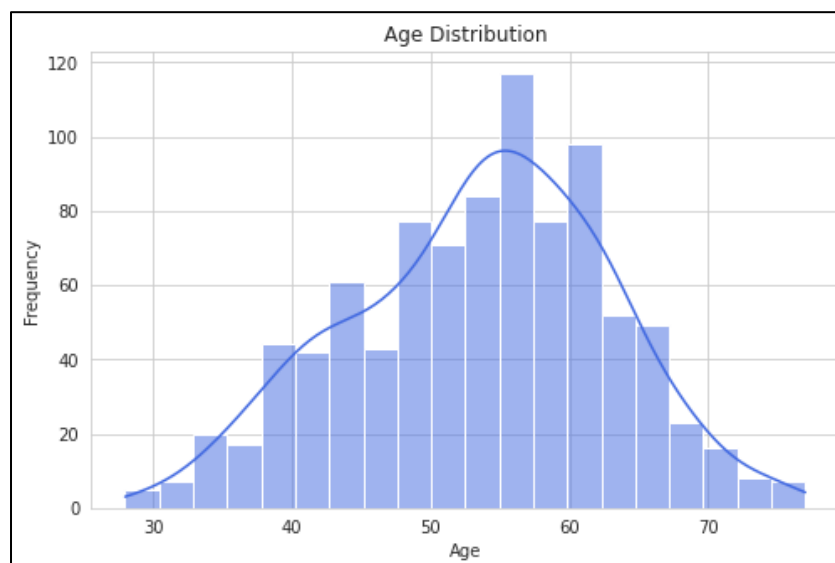
Key observations from the boxplot:

1. **Age :**
 - Relatively narrow distribution between approximately 30-80 years
 - Median around 55 years
 - Few outliers
2. **RestingBP:**
 - Ranges primarily between 100-175 mmHg
 - Median around 130 mmHg
 - Some lower outliers (including a value near 0, which is likely an error in the data)
3. **Cholesterol:**
 - Widest distribution of all features
 - Ranges primarily between 100-275 mg/dL
 - Median around 225 mg/dL
 - Multiple high outliers between 400-600 mg/dL
 - A few suspiciously low outliers near 0 (likely errors)
4. **MaxHR :**
 - Ranges primarily between 100-175 BPM
 - Median around 140 BPM
 - Some lower outliers

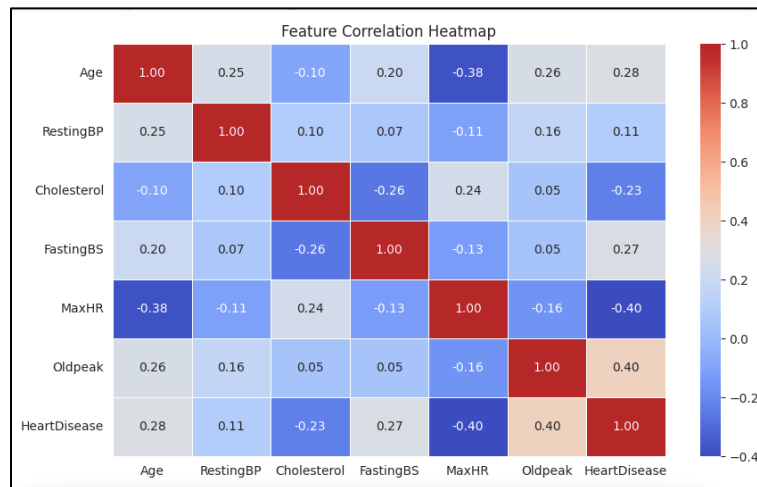
B. Pairwise Relationships of Features



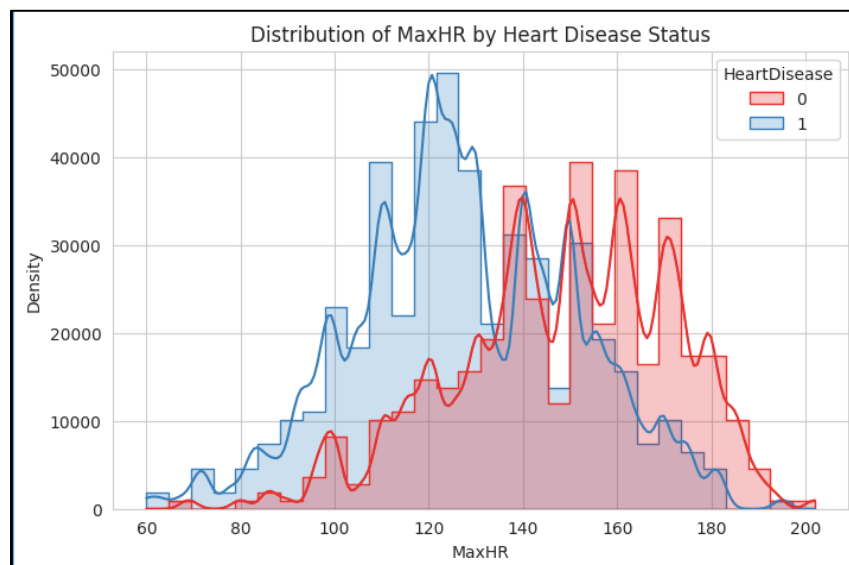
C. **Age Distribution:** Patients with heart disease tend to be older, with a higher prevalence after the age of 50.



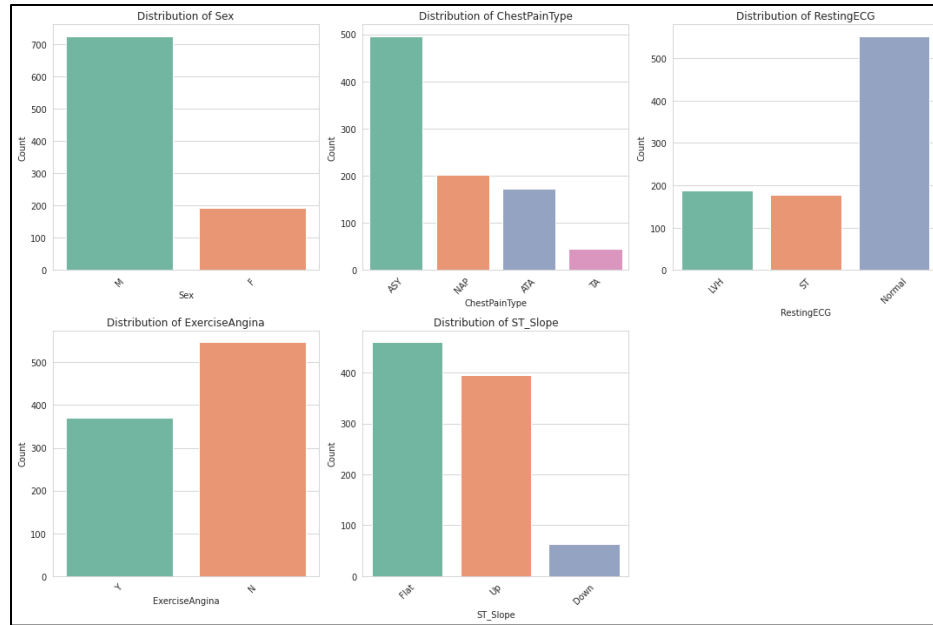
D. Correlation Heatmap: Oldpeak, ExerciseAngina, and ST_Slope have strong correlations with the presence of heart disease.



E. MaxHR Trends: Lower maximum heart rate is often associated with heart disease.

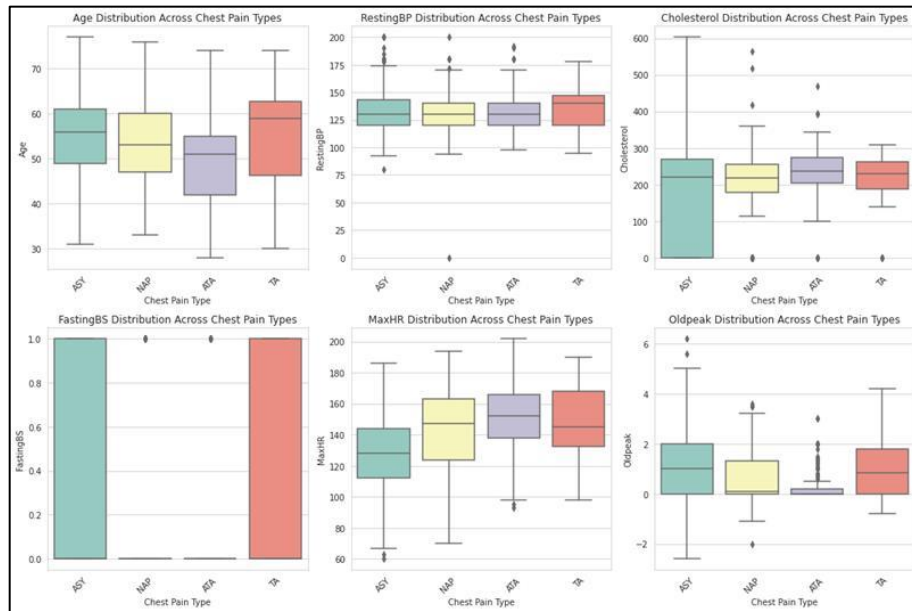


F. Distribution of Categorical Variables



- **Distribution by Sex:**
 - Shows a significant gender imbalance in the dataset
 - Males represent approximately 65-70% of the data
 - Females represent approximately 30-35% of the data
- **Distribution of ChestPainType:**
 - ASY is highest type of chest pain reported
 - Atypical angina and non-anginal pain have similar moderate frequencies
- **Distribution of RestingECG:**
 - Two categories with relatively even distribution
 - The blue bar (representing normal ECG) is higher than the others
- **Distribution of ExerciseAngina:**
 - Presence of angina is taller. Suggesting more patients experience angina during exercise than those who don't

G. Understanding relationship between categorical and numerical features

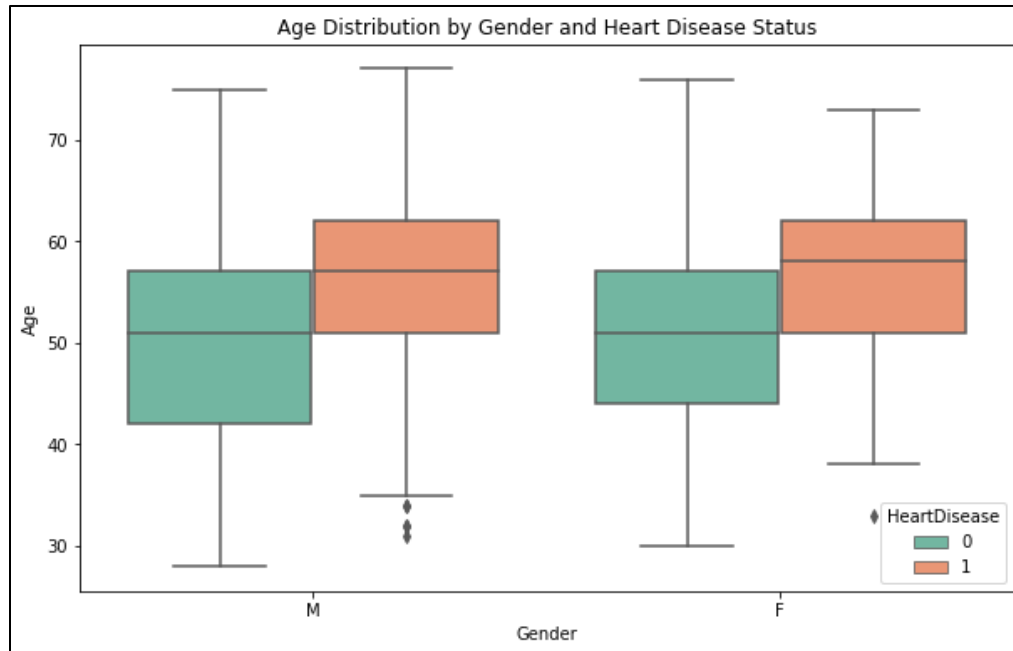


Key Takeaways from Each Plot

- Age Distribution Across CP Types
 - All CP types show a similar median age (~50–60 years).
 - Outliers exist for younger patients (~30s).
 - TA (Typical Angina) has a slightly higher median age compared to others.
- RestingBP Distribution Across CP Types
 - Median RestingBP is similar across groups (~120-140 mmHg).
 - ASY (Asymptomatic CP) shows higher outliers, indicating some patients with extremely high BP (~180-200 mmHg).
- Cholesterol Distribution Across CP Types
 - ASY has a wide spread, with some extreme outliers (~500-600 mg/dL).
 - Other CP types have a lower spread but similar median cholesterol levels (~200-300 mg/dL).
- FastingBS (Fasting Blood Sugar) Across CP Types
 - ASY and TA have higher FastingBS values (mostly 1s), meaning more patients in these groups have FBS > 120 mg/dL.
 - NAP and ATA have mostly 0s (lower FBS).
- MaxHR Distribution Across CP Types
 - ASY has the lowest median MaxHR (~120 bpm).
 - Other CP types have higher MaxHR (~140-160 bpm).
 - Outliers suggest some patients have unusually low MaxHR (~60-80 bpm).
- Oldpeak (ST Depression) Across CP Types
 - ASY has the highest Oldpeak values (~2-4), indicating higher ST depression, a potential marker for heart disease.
 - Other CP types have lower Oldpeak values.

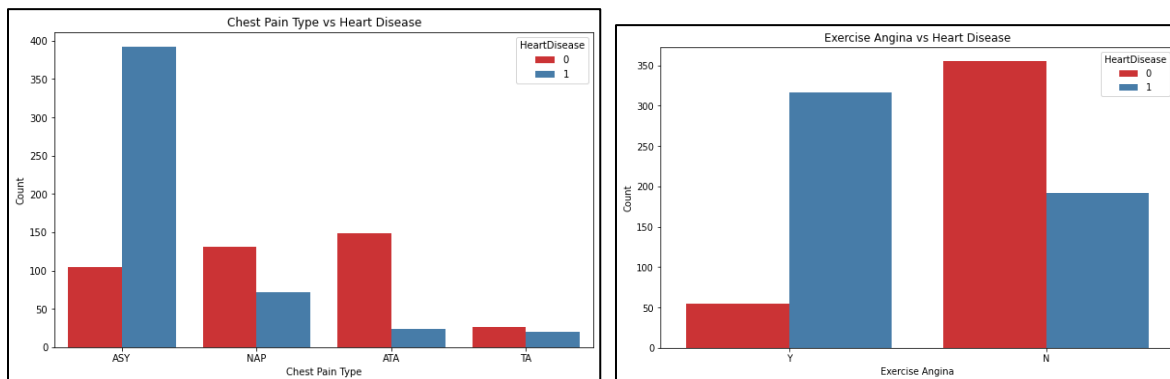
- Outliers show some negative values, which might be due to measurement errors or unique patient cases.

H. How does age influence the likelihood of developing heart disease across different gender groups?



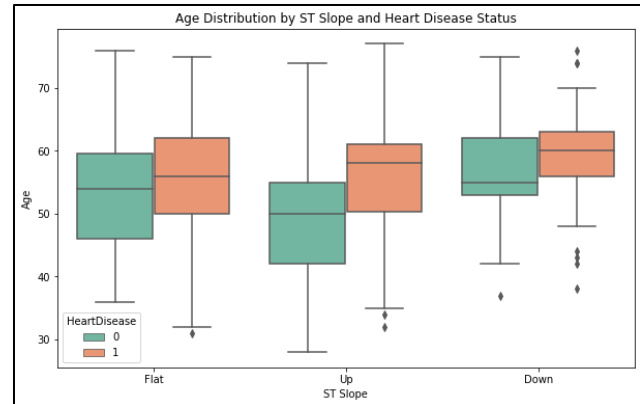
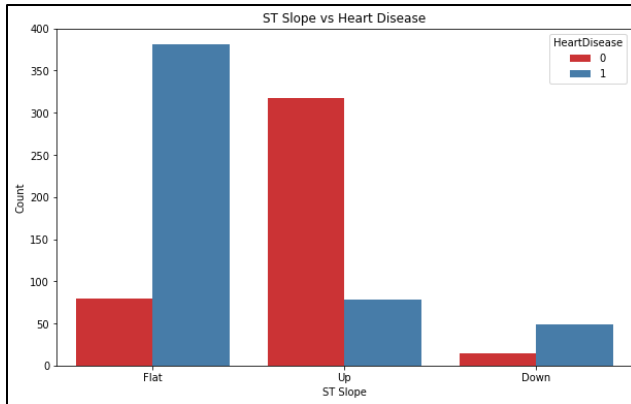
It is clear that as the age increases the likelihood of developing a heart disease increases.

I. Does the presence of chest pain and exercise-induced angina increase the risk of heart disease, and how do these factors interact with other variables?



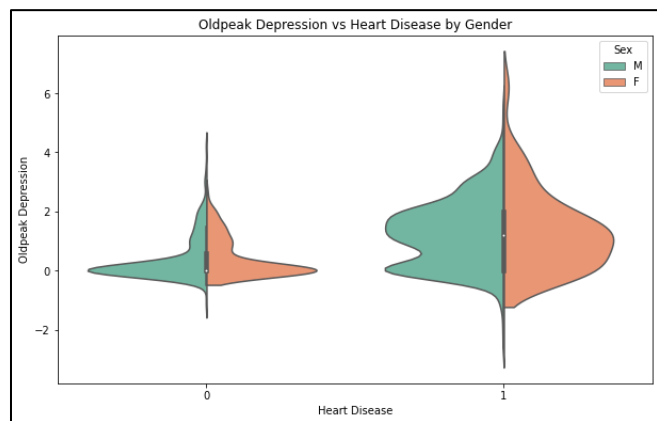
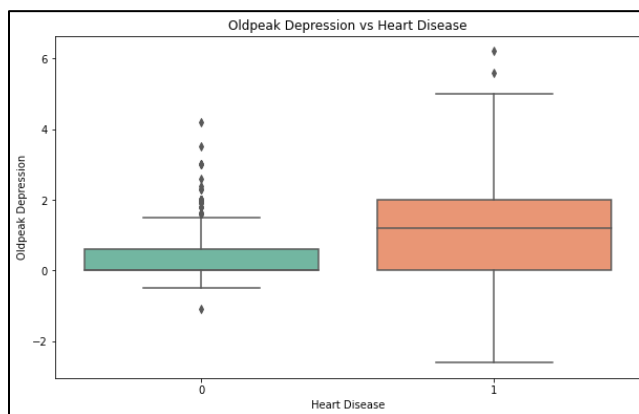
In case of ASY type of chest pain there is an extremely high chance of developing heart disease. Similarly, if there is pain in angina after exercising then there is a high chance of having heart disease.

J. Is there a significant difference in heart disease prevalence among individuals with varying ST slope levels, and how does it compare across different age groups?



A large number of individuals with a flat and downward ST slope have heart disease. The majority of individuals with an upward-sloping ST segment do not have heart disease. For each ST slope category, the age distribution of individuals with heart disease tends to be slightly higher than those without heart disease.

K. What patterns exist between old peak depression (from exercise) and heart disease occurrence, and do these patterns differ based on other clinical features?



The median Oldpeak value is higher for patients with heart disease. There is a wider **spread** in the Oldpeak values for those with heart disease, with several **outliers** at the higher end. Patients without heart disease (0) have Oldpeak values mostly concentrated around 0, with fewer extreme values. The distribution varies slightly between **males and females**, with **females exhibiting more extreme values**.

3.3 Key insights impacting business decisions.

1. Patient Demographics and Symptom Patterns

- **Significant gender imbalance:** Males (65-70%) dominate the patient population, suggesting targeted screening programs for men may yield higher detection rates
- **Asymptomatic chest pain (ASY)** is the most common presentation, highlighting the importance of proactive screening beyond symptom-based assessments
- **Exercise-induced angina** is prevalent among patients, indicating stress testing remains a valuable diagnostic tool

2. Clinical Risk Markers

- **Asymptomatic patients show concerning markers:** Higher blood pressure outliers (180-200 mmHg), extreme cholesterol levels (500-600 mg/dL), elevated fasting blood sugar, and significant ST depression
- **Asymptomatic patients have lower maximum heart rates** (~120 bpm vs. 140-160 bpm in other groups), potentially indicating compromised cardiac function
- **Normal ECG results** are common despite other risk factors, reinforcing that ECG alone may miss significant cardiac pathology

3. Age and Risk Factor Correlations

- **Similar age distribution across all chest pain types** (median 50-60 years), suggesting risk factors affect patients across similar age ranges regardless of symptom presentation
- **Typical Angina patients skew slightly older**, potentially representing a more advanced disease progression group

4. Strategic Implications

- Develop stronger screening protocols for asymptomatic patients, particularly males
- Emphasize multi-marker assessment approach rather than relying on single indicators like ECG or symptoms
- Consider lower thresholds for stress testing in high-risk demographic groups
- Incorporate maximum heart rate as a potential early warning indicator alongside traditional risk factors

5. Age as a Key Risk Factor

- The likelihood of developing heart disease increases with age, suggesting that **preventive screening and early lifestyle interventions** should be targeted toward **middle-aged and older individuals**.
- Business Decision: **Insurance companies** can adjust their health risk assessments based on age-related risk factors. **Healthcare providers** can focus more on heart disease awareness and prevention programs for older adults.

6. Oldpeak Depression as a Diagnostic Indicator

- Higher **Oldpeak depression values** (post-exercise ST depression) correlate with increased heart disease risk, with **females showing more extreme values** than males.
- Business Decision: **AI-driven diagnostics in wearable fitness devices** (like Apple Watch, Fitbit) and **hospital ECG screenings** can flag high Oldpeak values as a red flag for further cardiovascular testing.

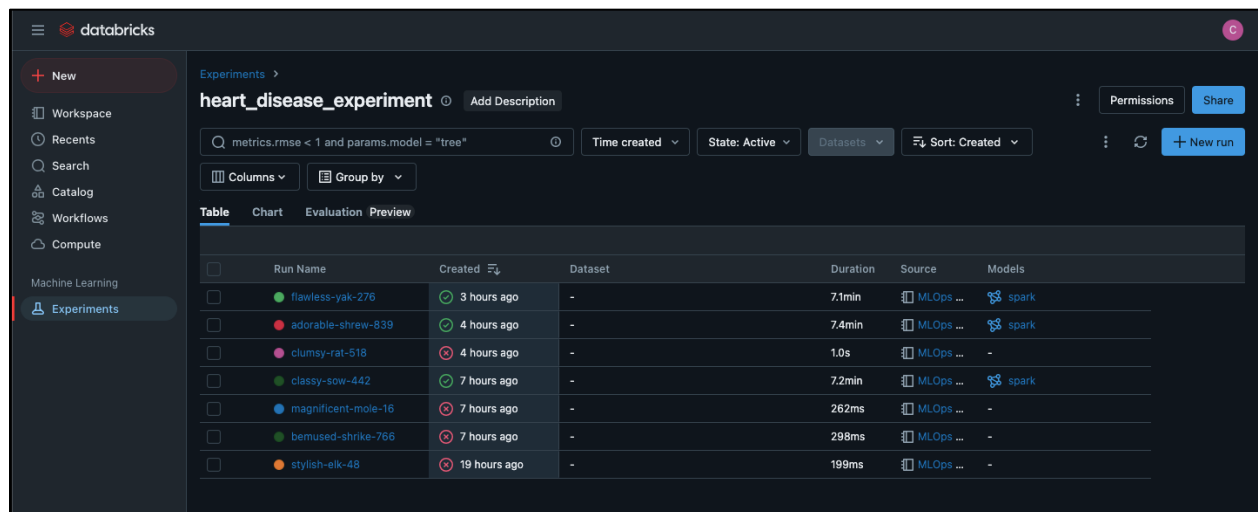
7. Predictive Modeling using Spark MLlib

5.1. Random Forest Model:

- The Random Forest algorithm was used for classification due to its robustness and ability to handle complex interactions between features.
- Parameters used:
 - a. Number of Trees: 100 (controls the ensemble size)
 - b. Max Depth: 10 (limits tree depth to avoid overfitting)
 - c. Min Instances per Node: 2 (ensures sufficient data per split)
 - d. Feature Subsampling Ratio: 0.8 (randomly selects features for training each tree)

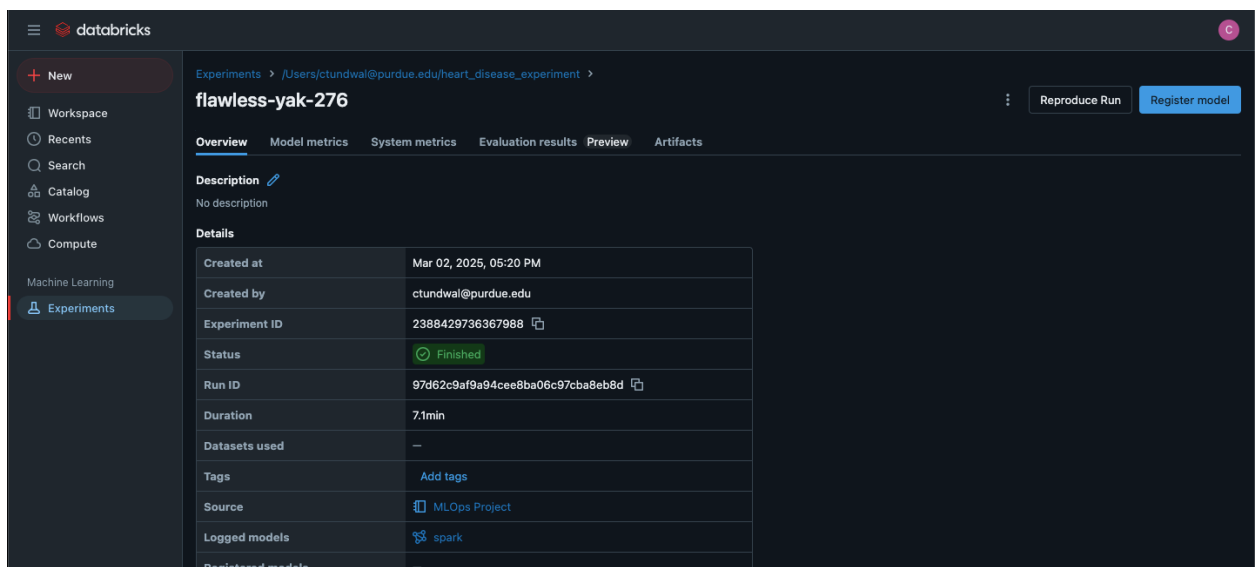
- The model was trained on the Spark MLlib framework, which efficiently processes large datasets using distributed computing.
- Hyperparameter Tuning: Cross-validation was applied to optimize hyperparameters such as learning rate, number of trees (for Random Forest).
- Model Evaluation Metrics:
 - a. Area Under the Curve (AUC) to evaluate classification performance.
 - b. Accuracy: The Random Forest model achieved an accuracy of 97.82%, demonstrating its effectiveness in correctly classifying heart disease presence.

8. Automation and Experiment Tracking



The screenshot shows the Databricks Experiments interface. The left sidebar contains navigation options: New, Workspace, Recents, Search, Catalog, Workflows, Compute, Machine Learning, and Experiments. The main panel displays the 'heart_disease_experiment' with a search filter 'metrics.rmse < 1 and params.model = "tree"'. A table lists several runs with their names, creation times, durations, and sources.

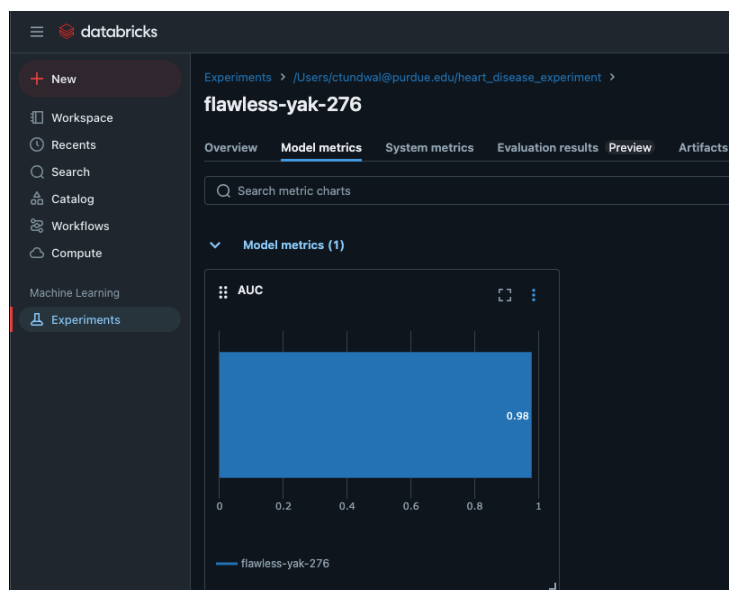
Run Name	Created	Dataset	Duration	Source	Models
flawless-yak-276	3 hours ago	-	7.1min	MLOps ...	spark
adorable-shrew-839	4 hours ago	-	7.4min	MLOps ...	spark
clumsy-rat-518	4 hours ago	-	1.0s	MLOps ...	-
classy-sow-442	7 hours ago	-	7.2min	MLOps ...	spark
magnificent-mole-16	7 hours ago	-	262ms	MLOps ...	-
bemused-shrike-766	7 hours ago	-	298ms	MLOps ...	-
stylish-elk-48	19 hours ago	-	199ms	MLOps ...	-



The screenshot shows the details for the 'flawless-yak-276' run. The left sidebar is the same as the previous screenshot. The main panel displays the 'flawless-yak-276' run with tabs for Overview, Model metrics, System metrics, Evaluation results, Preview, and Artifacts. The 'Overview' tab is selected, showing a description and a details table.

Details	
Created at	Mar 02, 2025, 05:20 PM
Created by	ctundwal@purdue.edu
Experiment ID	2388429736367988
Status	Finished
Run ID	97d62c9af9a94cee8ba06c97cba8eb8d
Duration	7.1min
Datasets used	-
Tags	Add tags
Source	MLOps Project
Logged models	spark
Registered models	-

Parameters (4)		Metrics (1)	
<input type="text" value="Search parameters"/>		<input type="text" value="Search metrics"/>	
Parameter	Value	Metric	Latest
best_impurity	gini	AUC	0.9782116959351276
best_maxDepth	10		
best_numTrees	50		
feature_columns	<div> ['Age', 'RestingBP', 'Cholesterol', 'FastingBS', 'MaxHR', 'Oldpeak', 'Sex_encoded', 'ChestPainType_encoded', 'RestingECG_encoded', 'ExerciseAngina_encoded', 'ST_Slope_encoded'] </div>		



5. Insights & Business Recommendations

- The model successfully identified key predictors of cardiovascular disease, aiding in risk assessment.
- Healthcare providers can use this model to prioritize high-risk patients for early intervention.
- Model performance visualization showcases how well predictions align with actual outcomes, validating its effectiveness in real-world applications.
- **Insurance Companies:** Implement dynamic, risk-based premium models using predictive heart disease risk factors.
- **Wearable Tech & AI Diagnostics:** Enhance ECG monitoring tools with ST slope and Oldpeak depression analysis.
- **Public Health Initiatives:** Focus awareness campaigns on older adults and high-risk groups to promote early prevention.
- Future improvements could integrate additional clinical variables and leverage advanced deep-learning techniques for enhanced accuracy.