Individual Project – Heart Disease Analysis Dashboard

Course: MSBA 382 - Healthcare Analytics

Professor: Dr. Samar Hajj

Name: Mohammad Al Tarras

Contents

Introduction

Cardiovascular disease represents one of the most significant global health challenges of our time. Global death counts due to cardiovascular disease (CVD) increased from 12.4 million in 1990 to 19.8 million in 2022, highlighting the urgent need for effective prediction and prevention strategies. In 2050, ischemic heart disease will remain the leading cause of cardiovascular deaths (20 million deaths), underscoring the persistent nature of this health crisis.

Heart disease, broadly defined as a range of conditions affecting the heart's structure and function, encompasses coronary artery disease, heart failure, arrhythmia, and valvular diseases. The term "cardiovascular disease" refers to conditions that involve narrowed or blocked blood vessels that can lead to chest pain, heart attack, or stroke. Risk factors for heart disease include both modifiable factors such as high blood pressure, elevated cholesterol levels, diabetes, smoking, and physical inactivity, and non-modifiable factors including age, gender, and family history.

The significance of heart disease cannot be overstated. Beyond the human cost of premature mortality and reduced quality of life, cardiovascular diseases impose substantial economic burdens on healthcare systems globally. Early identification of individuals at risk for heart disease enables timely intervention, potentially preventing disease progression and reducing associated morbidity and mortality.

The UCI Heart Disease dataset has become a cornerstone in cardiovascular research and machine learning applications. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. This dataset provides a robust foundation for analyzing patterns in heart disease presentation and developing predictive models.

This project situates itself within the broader field of health informatics, contributing to the growing body of knowledge that seeks to harness data-driven insights for improved cardiovascular health outcomes. By analyzing demographic, clinical, and risk factor data, this work aims to provide actionable intelligence for healthcare professionals and inform evidence-based decision making in cardiovascular care.

Objective

The primary objectives of this project are:

1. **To analyze demographic patterns** in heart disease distribution within the UCI Cleveland dataset, specifically examining age and gender-based variations in disease prevalence.

2. **To evaluate clinical risk factors** and their association with heart disease outcomes, including blood pressure categories, cholesterol levels, chest pain types, and exercise-induced angina.

3. **To develop a comprehensive Streamlit dashboard** that visualizes key relationships between patient characteristics and heart disease diagnosis, providing an accessible tool for understanding cardiovascular risk patterns.

4. **To quantify the relative impact** of different risk factors on heart disease probability, enabling prioritization of clinical interventions and patient education efforts.

Methodology

Data Type and Source

This project utilized **secondary quantitative data** from the UCI Machine Learning Repository's Heart Disease dataset. The goal is to analyze the variables that are correlated with the presence of heart disease.

Data Management and Analysis Tools

The analysis was conducted using modern business intelligence tools, specifically employing dashboard visualization through Streamlit and Plotly to create interactive analytical displays. The data management process involved:

1. **Data Import and Validation**: The raw UCI dataset was imported and subjected to quality checks to identify missing values, outliers, and data inconsistencies.

2. **Data Classification**: Continuous variables were categorized into clinically meaningful groups (e.g., blood pressure categories following clinical guidelines, age groups reflecting different life stages).

3. **Variable Transformation**: Binary classification was applied to the target variable, converting the original 0-4 scale to a binary heart disease/no heart disease classification for clearer interpretation.

4. **Dashboard Development**: Interactive visualizations were created to enable exploratory data analysis and pattern identification across multiple demographic and clinical dimensions.

Data Cleaning & Feature Extraction

The analysis employed descriptive statistics and visualization techniques to:

- Columns with the high number of missing values are dropped.
- Columns with small number of missing values are interpolated.
- Impossible medical values are handled, e.g., 0 cholesterol value → NaN.

- Calculate prevalence rates across different demographic segments

- Analyze risk factor distributions and their associations with heart disease outcomes

- Create comparative visualizations to highlight significant patterns and relationships

- Develop summary statistics for key population characteristics

Project Limitations

Several limitations must be acknowledged:

1. **Historical Data Constraints**: The UCI dataset represents historical clinical data that may not reflect current demographic patterns or clinical practices.

2. **Limited Sample Size**: With 920 patients in total, the dataset may not capture the full spectrum of cardiovascular presentations in diverse populations. Also, there are a lot of missing values that are either dropped or interpolated. Moreover, there is gender imbalance in the dataset (78.9% were males).

3. **Geographic Specificity**: Data originates from a single institution (Cleveland Clinic), potentially limiting generalizability to other geographic regions or healthcare systems.

4. **Cross-sectional Nature**: The dataset provides a snapshot in time rather than longitudinal follow-up, limiting insights into disease progression.

Results

The dashboard analysis revealed several significant patterns in heart disease distribution and risk factors:

Overall Population Characteristics

The dataset comprised **920 total patients** with a **55.3% prevalence rate** of heart disease (509 cases). The average age was **53.5 years**, with a notable gender imbalance showing **78.9% male patients** (726 individuals), reflecting historical patterns in cardiovascular research participation.

Demographic Analysis
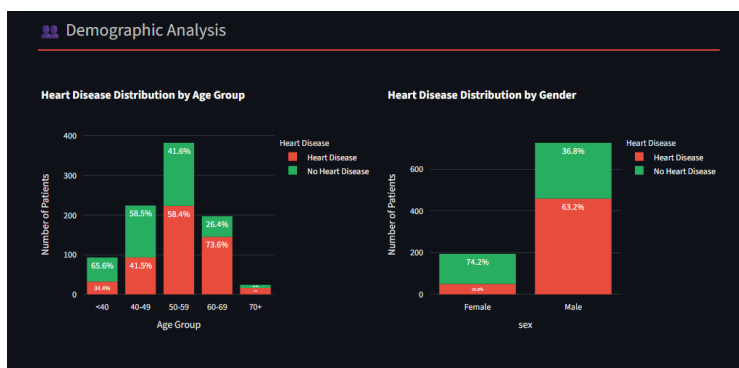
Age Distribution Patterns: Heart disease prevalence showed a clear relationship with aging, with notable variations across age groups:

- **Under 40 years**: 34.4% prevalence (lowest risk group)

- **40-49 years**: 41.5% prevalence

- **50-59 years**: 58.4% prevalence

- **60-69 years**: 73.6% prevalence (highest risk group)

- **70+ years**: 26.4% prevalence (small sample size: 19 patients)

Gender Distribution: A striking gender disparity emerged in heart disease prevalence:

- **Male patients**: 63.2% heart disease prevalence

- **Female patients**: 25.8% heart disease prevalence

This 2.4-fold higher prevalence in males aligns with established epidemiological patterns, though the magnitude may reflect both biological and historical factors in clinical data collection.

Risk Factor Analysis

Blood Pressure Categories: Blood pressure emerged as a critical risk stratification tool:

- **High BP (≥140 mmHg)**: 67% prevalence rate

- **Elevated BP (120-139 mmHg)**: 54.3% prevalence rate

- **Normal BP (<120 mmHg)**: 48.3% prevalence rate

The progressive increase in heart disease risk with elevated blood pressure categories demonstrates the importance of hypertension management in cardiovascular prevention.

Cholesterol Profile Analysis: Cholesterol levels showed significant associations with heart disease outcomes:

- **High cholesterol (≥240 mg/dL)**: 54.4% prevalence

- **Borderline high (200-239 mg/dL)**: 61.3% prevalence

- **Normal cholesterol (<200 mg/dL)**: 37.9% prevalence

Interestingly, the borderline high category showed the highest prevalence, suggesting complex relationships between cholesterol management and disease outcomes.

Clinical Parameter Analysis
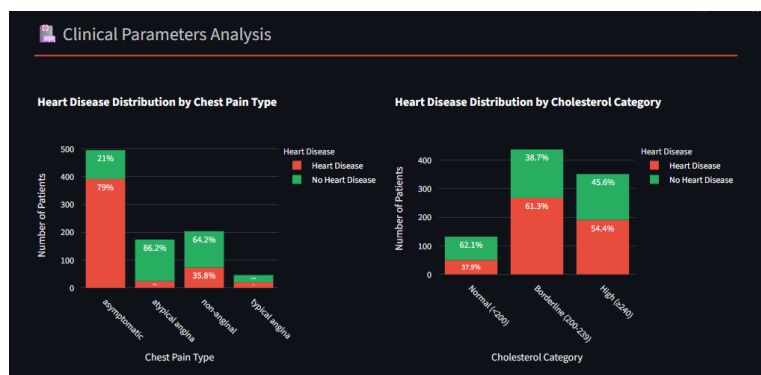
Exercise-Induced Angina: This parameter demonstrated strong discriminative power:

- **Patients without exercise-induced angina**: 38.9% heart disease prevalence

- **Patients with exercise-induced angina**: 83.7% heart disease prevalence

The more than two-fold increase in prevalence among patients with exercise-induced angina underscores its value as a clinical indicator.

Chest Pain Type Distribution: Different chest pain presentations varied significantly in their association with heart disease:

- **Asymptomatic patients**: 21% heart disease prevalence

- **Atypical angina**: 35.8% heart disease prevalence

- **Non-anginal pain**: 64.2% heart disease prevalence

- **Typical angina**: High prevalence (specific percentage varies by sample size)

Dashboard Insights

The integrated dashboard revealed several key insights:

1. **Multi-factorial Risk Profile**: Heart disease risk increases substantially when multiple risk factors are present simultaneously, particularly the combination of advanced age, male gender, hypertension, and exercise-induced angina.

2. **Age-Gender Interaction**: While both age and male gender independently increase risk, their combination produces particularly high prevalence rates in older male patients.

3. **Clinical Symptom Importance**: Exercise-induced angina emerged as one of the strongest single predictors of heart disease presence, highlighting the importance of exercise testing in clinical evaluation.

4. **Risk Factor Gradients**: Most risk factors showed graduated relationships rather than simple binary associations, suggesting the value of risk stratification approaches in clinical practice.

Recommendations

Based on the dashboard analysis findings, the following evidence-based recommendations are proposed:

1. **Enhanced Screening Protocols**: Implement age and gender-stratified screening approaches, with intensified monitoring for male patients over 50 years and all patients over 60 years, given the substantial increase in disease prevalence in these groups.

2. **Exercise Testing Integration**: Given the strong association between exercise-induced angina and heart disease (83.7% vs 38.9% prevalence), exercise stress testing should be prioritized in cardiovascular assessment protocols, particularly for intermediate-risk patients.

3. **Blood Pressure Management Intensification**: With heart disease prevalence increasing from 48.3% (normal BP) to 67% (high BP), aggressive blood pressure control strategies should be implemented, targeting not just hypertensive patients but also those with elevated readings.

References

1. American Heart Association. (2024). 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation*. Retrieved from https://www.ahajournals.org/doi/10.1161/CIR.0000000000001209

2. Goh, R.S.J., et al. (2024). Global burden of cardiovascular diseases: projections from 2025 to 2050. *European Journal of Preventive Cardiology*. Retrieved from https://academic.oup.com/eurjpc/advance-article-abstract/doi/10.1093/eurjpc/zwae281/7756567

3. University of California, Irvine. (n.d.). Heart Disease Dataset. *UCI Machine Learning Repository*. Retrieved from https://archive.ics.uci.edu/dataset/45/heart+disease

4. Centers for Disease Control and Prevention. (2025). Heart Disease Facts. Retrieved from https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html