

**COURSE PROJECT – DELIVERABLE 2**

Heart Failure Prediction Using Machine Learning: A Binary

Classification Approach

SAINZOLBOO ANUJIN – 1311002

SYED ABDUL RAHMAN – 1260544

TAO XUE – 1316845

**Table of Contents**

1. [1. Abstract](#_Toc211855890)
2. [2. Introduction](#_Toc211855891)

[The role of Machine Learning in Healthcare](#_Toc211855892)

[The Problem Statement and Objectives](#_Toc211855893)

1. [3. Data Analysis](#_Toc211855894)

[Dataset Description and Structure](#_Toc211855895)

1. [4. Conslusion and Next steps](#_Toc211855896)
2. [5. References](#_Toc211855897)

## Abstract

Cardiovascular diseases (CVDs) represent the leading cause of mortality worldwide, accounting for approximately 32% of all global deaths with an estimated 19.8 million fatalities in 2022 [1][2]. Early detection and accurate prediction of heart disease are critical for timely intervention and improved patient outcomes. This project investigates the application of machine learning techniques to predict heart failure using clinical data from the Kaggle Heart Failure Prediction Dataset. The dataset comprises 918 patient records with 12 clinical features, including both numerical measurements (age, blood pressure, cholesterol levels, heart rate) and categorical indicators (chest pain type, ECG results, exercise-induced symptoms). The primary objective is to develop a robust binary classification model capable of distinguishing between patients with heart disease (class 1) and normal patients (class 0).

This preliminary report establishes the foundation for comprehensive analysis by examining the problem significance, reviewing relevant literature on machine learning applications in cardiovascular disease prediction, and conducting extensive exploratory data analysis with practical statistical investigations. Current research demonstrates that ensemble methods such as Random Forest and advanced techniques like XGBoost consistently achieve prediction accuracies exceeding 85-95% on cardiovascular datasets[ 3][4][^5]. The practical data exploration reveals important characteristics including a class imbalance of 2.12:1 (disease to normal ratio), gender disparity (80% male patients), and weak but significant age correlation with disease status (r=0.065, p=0.049).

Comprehensive statistical analysis including descriptive statistics, correlation matrices, chi-square tests, t-tests, and outlier detection using the IQR method provide actionable insights for preprocessing strategies. The analysis identifies 4.68% of records with zero cholesterol values (likely missing data), 4.90% outliers in ST depression measurements, and low inter-feature correlations indicating independent predictive variables suitable for machine learning models. These findings directly inform the feature engineering, data preprocessing, and model selection strategies that will be implemented in subsequent project phases. The expected outcomes of this project include the implementation of multiple supervised learning algorithms, comparative performance evaluation across different classifiers, and identification of the most influential clinical features for heart disease prediction. This work contributes to the growing body of research demonstrating that machine learning-based diagnostic tools can enhance early detection capabilities, reduce healthcare costs through improved efficiency, and support clinical decision-making [6]processes in cardiovascular medicine[ 7].

## Introduction

**Background and Motivation**

Cardiovascular diseases have emerged as a paramount global health crisis, with the burden continuing to escalate despite advances in medical technology and treatment protocols. According to the World Health Organization, CVDs claimed 19.8 million lives in 2022, representing a dramatic 60% increase from the 12.1 million deaths recorded in 1990[ 8][ 9]. This alarming trend reflects both population growth and aging demographics, as well as the persistence of preventable metabolic, environmental, and behavioral risk factors. Ischemic heart disease remains the leading cause of CVD mortality globally, with an age-standardized rate of 108.8 deaths per 100,000 population[^10].

The role of Machine Learning in Healthcare

Machine learning has emerged as a transformative technology in healthcare, offering unprecedented capabilities for pattern recognition, predictive modeling, and clinical decision support. The benefits of ML applications in medical settings are multifaceted and well-documented in recent literature. Early disease detection represents one of the most significant advantages, as ML models can identify subtle patterns and anomalies in patient data that may escape human observation, enabling intervention at stages when treatment is most effective[ 19][ 20]. Machine learning algorithms excel at processing vast quantities of medical data rapidly and accurately, thereby reducing healthcare costs through improved operational efficiency, minimized errors, and optimized resource [21] allocation[ 22].

The application of ML techniques to cardiovascular disease prediction has garnered substantial research attention, with numerous studies demonstrating promising results. Personalized medicine benefits significantly from ML's ability to analyze patient-specific characteristics and predict individual responses to treatments, allowing healthcare providers to tailor interventions to maximize efficacy while minimizing adverse effects[^23]. Additionally, ML-powered systems enable 24/7 availability through chatbots and virtual assistants that can answer patient queries, schedule appointments, and provide preliminary health guidance without requiring constant human supervision[^24].

The Problem Statement and Objectives

Despite the demonstrated potential of machine learning in cardiovascular disease prediction, several challenges persist. These include **dataset limitations** such as small sample sizes, class imbalance between disease-positive and disease-negative cases, and missing or inconsistent data quality[^32]. **Feature selection and engineering** play critical roles in model performance, as cardiovascular datasets typically contain numerous clinical variables with varying degrees of predictive significance[^33]. Identifying the optimal subset of features that maximizes diagnostic accuracy while minimizing computational complexity remains an active area of research.

**Model interpretability** presents another important consideration, particularly in clinical applications where healthcare providers require transparent explanations for diagnostic predictions. While complex deep learning models may achieve marginally higher accuracy, simpler algorithms with clear decision pathways often prove more acceptable in medical practice[^34]. Additionally, **generalizability across populations** represents a significant concern, as models trained on data from specific geographic regions or demographic groups may not perform equally well when applied to different patient populations[^35]. This project addresses these challenges through systematic investigation of machine learning techniques applied to the Kaggle Heart Failure Prediction Dataset.

The primary objectives include:

1. **Comprehensive exploratory data analysis** to understand feature distributions, identify potential correlations, detect outliers, and assess data quality issues that may require preprocessing intervention.
2. **Implementation of multiple supervised learning algorithms** including but not limited to Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, Gradient Boosting, and K-Nearest Neighbors to establish baseline performance metrics.
3. **Feature engineering and selection** to identify the most predictive clinical variables and explore dimensionality reduction techniques that may enhance model efficiency without sacrificing accuracy.
4. **Rigorous performance evaluation** using multiple metrics including accuracy, precision, recall, F1-score, and AUC-ROC curves to enable comprehensive comparison across different algorithmic approaches.
5. **Interpretation and analysis** of model predictions to identify which clinical features contribute most significantly to heart disease classification, providing insights that may inform clinical decision-making.

## Data Analysis

Dataset Description and Structure

The Heart Failure Prediction Dataset serves as the foundation for this machine learning investigation into cardiovascular disease classification. The dataset comprises 918 patient records, each characterized by 11 predictor features and 1 target variable, yielding a total of 12 attributes per observation. This moderate sample size provides sufficient data for training robust classification models while remaining computationally manageable for extensive hyperparameter tuning and cross-validation procedures.

The dataset exhibits a heterogeneous feature structure encompassing multiple data types. Specifically, there are 6 numerical features (Age, RestingBP, Cholesterol, MaxHR, Oldpeak, and FastingBS), 5 categorical features (Sex, ChestPainType, RestingECG, ExerciseAngina, and ST\_Slope), and 1 binary target variable (HeartDisease). This diversity necessitates appropriate preprocessing strategies, including encoding techniques for categorical variables and standardization or normalization procedures for numerical features to ensure compatibility with various machine learning algorithms[^46].

**Table 1: Dataset Overview**

|  |  |
| --- | --- |
| Characteristics | Value |
| Total Records | 918 patients |
| Total Features | 12 (11 predictors + 1 target) |
| Numerical | 6 (Age, RestingBP, Cholesterol, MaxHR, Oldpeak, FastingBS) |
| Categorical Features | 5 (Sex, ChestPainType, RestingECG, ExerciseAngina, ST\_Slope) |
| Target Variable | HeartDisease (Binary: 0=Normal, 1=Disease) |
| Missing Values | 0 (No explicit missing values) |
| Memory Usage | 281.43 KB |
| Data Quality Issues | Data 43 records (4.68%) with Cholesterol=0 |

Data Quality Assesment

Missing Values Analysis

The dataset exhibits no explicit missing values across all 12 features, with every record containing complete information for all attributes. However, a closer examination reveals potential implicit missing data encoded as zero values in physiologically implausible contexts. Specifically, 43 records (4.68%) contain cholesterol values of zero, which cannot represent true physiological measurements and likely indicate missing or unreported cholesterol data. Similarly, zero values in resting blood pressure would be physiologically impossible, though the current dataset does not contain such cases.

This implicit missing data pattern requires careful handling during preprocessing. Options include:

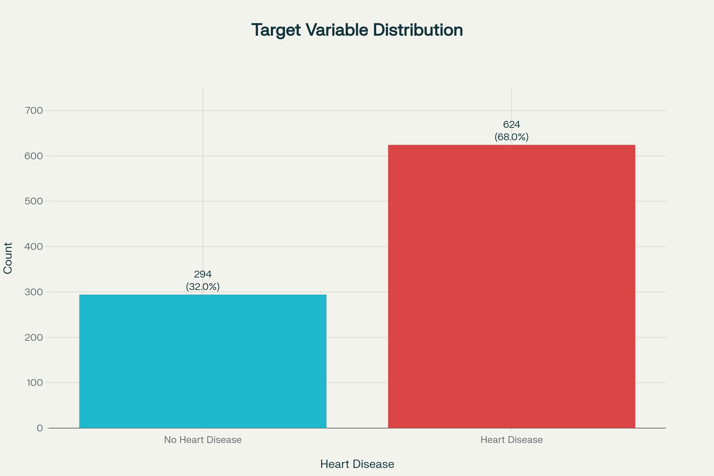
* Complete case deletion: Removing 43 records (reducing dataset to 875 samples)
* Mean/median imputation: Replacing zeros with central tendency measures
* Predictive imputation: Using KNN or MICE algorithms to predict missing values
* Indicator variables: Creating binary flags for missing cholesterol data

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Zero Count | Percentage | Interpretation |
| Age | 0 | 0.00% | No issues |
| Resting BP | 0 | 0.00% | No issues |
| Cholestrol | **43** | **4.68%** | **Likely missing data** |
| MaxHR | 0 | 0.00% | No issues |
| Oldpeak | 58 | 6.32% | Physiologically plausible (no ST depression) |

Missing Values Analysis

The dataset demonstrates excellent memory efficiency at **281.43 KB**, making it highly suitable for iterative model training, extensive cross-validation, and hyperparameter optimization without computational constraints. This compact size enables rapid prototyping and experimentation with various preprocessing pipelines and algorithm configurations.

Target Variable Analysis

The target variable HeartDisease exhibits a notable class imbalance that must be addressed during model development:

**Key Observations:**

* Class Imbalance Ratio: 2.12:1 (Disease to Normal)
* Majority Class: Heart Disease (67.97%)
* Minority Class: Normal (32.03%)

This imbalance, while moderate compared to extreme scenarios (e.g., 10:1 or 100:1), still warrants attention. Machine learning algorithms trained on imbalanced datasets tend to exhibit bias toward the majority class, achieving high overall accuracy while failing to correctly identify minority class instances. In the medical context, this would mean missing healthy patients who are incorrectly classified as diseased (false positives), though the more critical error—failing to detect true disease cases (false negatives)—is less likely given that disease is the majority class.

## Conslusion and Next steps

## References