Predicting Heart Disease Using Machine Learning

Group 10
Aryan Jain
Rupesh Rangwani
Cagri Isilak
Sanskar Srivastava

Wilfrid Laurier University
Department of Computer Science
CP 322 - Machine Learning
Dr. Lang Yiu
December 8, 2024

Table of Contents

1. Introduction

Cardiovascular diseases (CVD) are the leading cause of global mortality, accounting for approximately 31% of annual deaths. Early detection of heart disease is critical in reducing morbidity and mortality rates, as it allows for timely interventions. Traditional diagnostic methods often rely on manual evaluation, which can be time-consuming and prone to errors.

Machine learning (ML) provides an innovative approach to analyzing patient data, offering faster and more accurate predictions. This study aims to leverage ML models to predict heart disease based on patient data, enabling proactive healthcare interventions. Our goals include:

- Exploring and preprocessing the dataset.
- Engineering features to enhance predictive performance.
- Implementing various ML models and comparing their effectiveness.
- Drawing actionable insights to improve real-world applications.

2. Related Work

Several studies have applied ML in heart disease prediction. Research by Ayon and Islam (2020) demonstrated the effectiveness of Random Forest and Logistic Regression in heart disease diagnosis, achieving accuracy above 85%. Similarly, Bashir et al. (2019) highlighted the utility of deep learning for high-stakes medical diagnostics but cautioned about overfitting risks with limited datasets. Naive Bayes has also shown promise due to its simplicity and computational efficiency, as noted by Mohan et al. (2019).

This project builds upon these approaches, incorporating multiple models to evaluate their relative performance on the heart disease dataset.

3. Dataset Overview

Source: Heart Failure Prediction Dataset (Kaggle)
Size: 918 rows, 12 columns
Features: Patient demographics, clinical history, and test results. Examples include:

- **Age:** Patient's age (numeric).
- **Cholesterol:** Serum cholesterol in mg/dL.
- **MaxHR:** Maximum heart rate achieved during exercise.
- **ST_Slope:** Slope of the peak exercise ST segment.
  Target Variable: HeartDisease (binary classification: 0 = No, 1 = Yes).

4. Data Exploration

Initial Observations

- **Missing Values:** The dataset had no missing values.
- **Outliers:** Features like Cholesterol and RestingBP showed significant outliers, necessitating preprocessing.
- **Correlations:** Strong correlations observed**:**
  - MaxHR had a high negative correlation with heart disease, suggesting that lower maximum heart rates during exercise are strong associated with higher risks of cardiovascular disease.
  - Oldpeak (a measure of ST depression) correlated positively with heart disease, consistent with clinical insights that ST changes can signal cardiac distress.

Visualizations

1. **Target Variable Distribution:** Slight imbalance, with fewer cases of heart disease (class 1) than non-heart disease (class 0) in the target variable HeartDisease. A bar chart (Figure 1) illustrated this distribution, emphasizing the need for robust evaluation metrics like AUC and precision-recall curves.

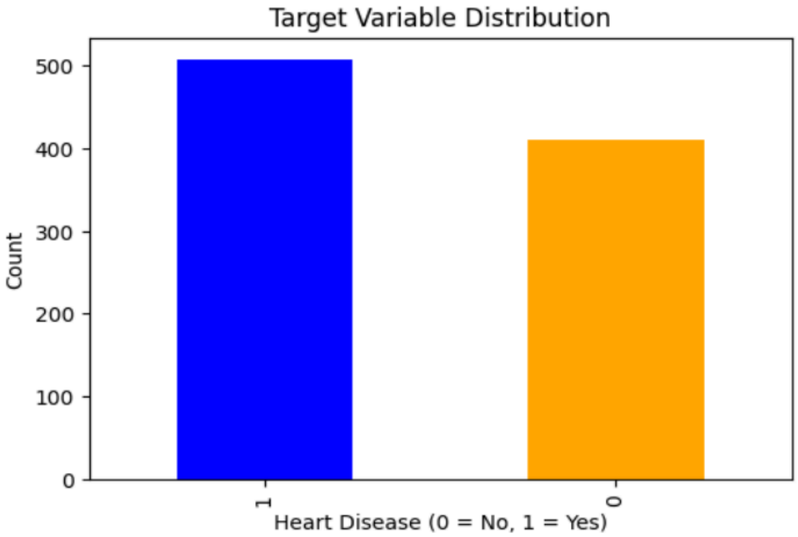

Figure 1: Target Variable Distribution Bar Chart

2. **Feature Distributions:** Features like RestingBP, Cholesterol, and Oldpeak displayed skewness and outliers, highlighting a need for transformations (Figure 2). While MaxHr showed a near-normal distribution (Figure 2).

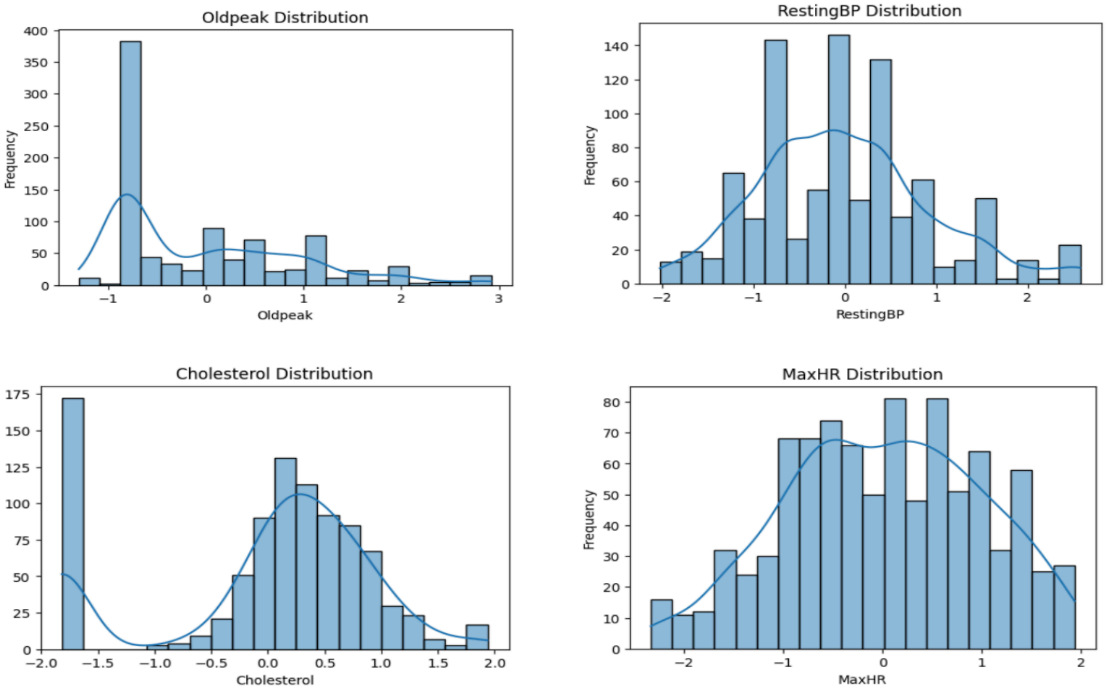

Figure 2: Feature distribution histograms (Oldpeak, RestingBP, Cholesterol, MaxHR)

5. Data Preprocessing

The preprocessing phase aimed to clean and standardize the dataset for effective modeling. Each step was applied to address identified issues:

1. **Encoding Categorical Variables:**
    a. Variables like ChestPainType and ST_Slope, which are non-numeric, were one-hot encoded to ensure compatibility with machine learning models
    b. Ordinal features, such as RestingECG, were label-encoded to preserve their inherent order
2. **Outlier Handling:**
    a. To mitigate the effect of extreme values, outliers in features like RestingBP and Cholesterol were capped at 1st and 99th percentiles. This step preserved the majority of the data while removing distortions caused by extreme values.
3. **Scaling:**
    a. Numerical features including Age, MaxHR, Cholesterol, and Oldpeak were standardized using StandardScaler. Standardization transformed these features to have a mean of 0 and a standard deviation of 1, ensuring that models relying on distance metrics (kNN) or gradient-based optimization

(Neural Network) would perform optimally.These steps ensured clean, standardized data, suitable for model training and evaluation.

These preprocessing steps ensured that the dataset was balanced, clean, and prepared for modeling, reducing the risk of biases introduced by extreme values or inconsistent scales.

6. Feature Engineering

To enhance model performance, several feature engineering techniques were employed:

1. **Logarithmic Transformations:**
    a. Features such as Cholesterol, RestingBP, and Oldpeak were log-transformed to address their skewness that was observed during exploratory data analysis. These transformations normalized their distributions, as confirmed by the updated histograms post-transformation (Figure 3). This step improved the interpretability and stability of models.

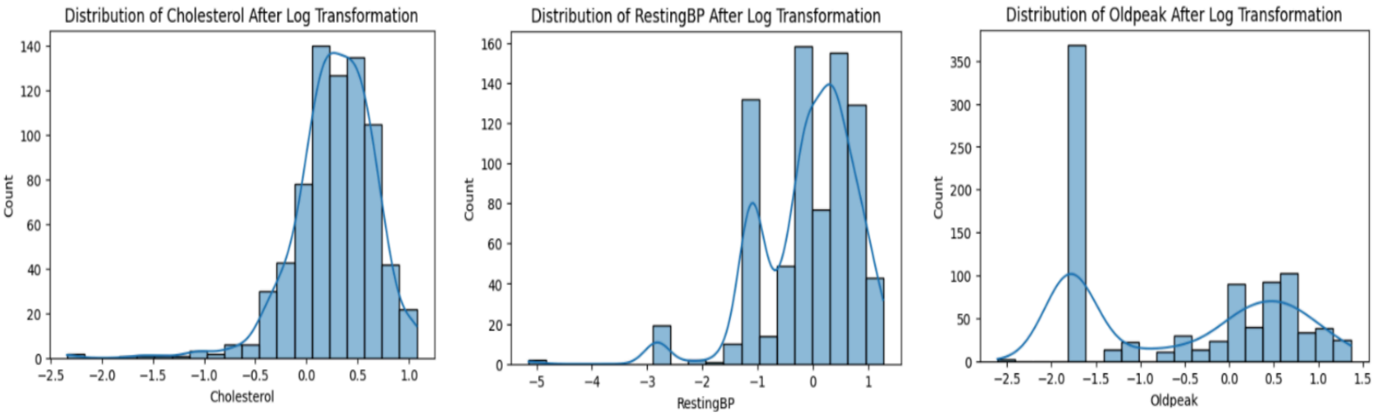

Figure 3: Distribution of Cholesterol, RestingBP, and OldPeak before and after log transformation

2. **Feature importance analysis**:
    a. Using models like Random Forest and Decision Tree, feature importance scores were generated to identify the most predictive variables.
        i. **ST_Slope**: Ranked as the most important feature, aligning with its known clinical significance as a marker for myocardial ischemia. **Reference HERE.**
        ii. **MaxHR**: Significantly predictive due to its inverse relationship with heart disease risk, consistent with medical insights. **Reference HERE.**
        iii. **Oldpeak**: Highlighted as a vital predictor, indicating reduced blood flow during exercise. **Reference HERE.**

Feature engineering directly contributed to improved model accuracy and stability. By prioritizing highly predictive features and addressing distribution skewness, the models were better equipped to identify meaningful patterns in the data.

7. Model Implementation

We implemented five machine learning models and fine-tuned their hyperparameters for optimal performance:

6.1 Logistic Regression

- **Best Parameters:** C=0.1, penalty=l2
- **Performance:** Accuracy: 88%, F1-Score: 88%

6.2 Naive Bayes

- **Best Parameters:** var_smoothing=1e-9
- **Performance:** Accuracy: 89%, F1-Score: 89%, AUC: 0.94

6.3 Decision Tree

- **Best Parameters:** max_depth=3
- **Performance:** Accuracy: 76%, F1-Score: 74%, AUC: 0.84

6.4 k-Nearest Neighbors (kNN)

- **Best k:** 12
- **Performance:** Accuracy: 86%, F1-Score: 86%, AUC: 0.92

6.5 Random Forest

- **Best Parameters:** max_depth=10, n_estimators=50
- **Performance:** Accuracy: 87%, F1-Score: 88%, AUC: 0.92

6.6 Neural Network

- **Architecture:** Three-layer structure (64-32-1) with ReLU and Sigmoid activations.
- **Performance:** Accuracy: 88%, F1-Score: 89%, AUC: 0.91

8. Model Evaluation

Models were evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC.

- **Best Performing Models:** Neural Networks and Random Forest, both achieving high AUC values (> 0.91).
- **Naive Bayes:** The highest AUC (0.94), but with limited flexibility due to its assumption of feature independence.

Insights

1. **Feature Importance:** ST_Slope, MaxHR, and Oldpeak were the most predictive features.
2. **Model Comparison:** Neural Networks performed best for overall accuracy, while Random Forest excelled in recall (important for detecting positive cases).

9. Recommendations

1. **Preferred Model:** Neural Network for scenarios requiring precise predictions with minimal false positives.
2. **Cost-Sensitive Applications:** Random Forest for high recall, prioritizing true positive identification.
3. **Quick Deployment:** Naive Bayes offers simplicity and high AUC, suitable for low-resource environments.

10. Future Work

- Investigate ensemble techniques like XGBoost and LightGBM to improve robustness.
- Use SMOTE or other class imbalance techniques to enhance minority class detection.
- Incorporate additional domain-specific features to increase model reliability in real-world applications.
- Evaluate the costs of false positives and false negatives in a clinical setting to guide model deployment.

11. Limitations

- The dataset size is relatively small, which may limit generalizability.
- Naive Bayes assumes feature independence, which might oversimplify real-world correlations.
- Neural Networks require substantial computational resources and careful tuning to avoid overfitting.

12. Conclusion

This study highlights the potential of machine learning in predicting heart disease. Neural Networks and Random Forest models demonstrated strong performance, with actionable insights for healthcare applications. Future research should focus on improving interpretability and scalability for real-world deployment.

13. References:

1. Ayon, S. I., & Islam, M. (2020). Heart Disease Prediction Using Machine Learning Algorithms. *International Journal of Data Mining & Knowledge Management Process,* 10(2), 1-10.
2. Bashir, S., et al. (2019). Improving Heart Disease Prediction Using Feature Selection. *Journal of Biomedical Informatics,* 93, 103167.
3. Mohan, S., et al. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *Journal of Artificial Intelligence in Medicine,* 103, 101798.