Heart Disease Prediction and Risk Analysis by Machine Learning Techniques

Omkar Singh[1], Amit Kumar Pandey[2], Ninad Sarang[3]

[1]HOD of Data Science, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai-400101, Maharashtra, India
[2]Assistant Professor, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai-400101, Maharashtra, India
[3]PG Student, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai-400101, Maharashtra, India

Abstract

Cardiovascular disease remains one of the leading causes of global mortality, underscoring the urgent need for robust early detection systems. This study introduces an innovative machine learning framework for heart disease prediction using the Cleveland dataset (sourced from the UCI repository via Kaggle). We employ a Decision Tree classifier for its transparency and visual interpretability, and benchmark its performance against a Gaussian Naïve Bayes model. Our data pipeline includes rigorous preprocessing steps—median imputation, Min-Max normalization, and stratified train-test partitioning—to ensure high-quality inputs. Visual outputs, including a detailed Decision Tree diagram (Figure 1) and a Principal Component Analysis (PCA) projection (Figure 2), elucidate the underlying data structure and decision boundaries. Although the Gaussian Naïve Bayes model demonstrates higher accuracy, the Decision Tree's clear decision paths offer invaluable insights for clinical applications. This work balances predictive performance with interpretability, paving the way for future research with hybrid models and expanded clinical feature sets.

Keywords: Heart disease prediction, machine learning, Decision Tree, Naïve Bayes, interpretability, risk analysis, Cleveland dataset, healthcare analytics.

I. Introduction

Cardiovascular diseases (CVDs) contribute to approximately 17.9 million deaths annually, making early diagnosis essential for reducing mortality and curtailing healthcare costs. The multifaceted nature of heart disease—driven by factors such as age, cholesterol, and blood pressure—poses significant challenges to early detection. Recent breakthroughs in machine learning have unlocked new avenues for analyzing complex clinical data, offering promising alternatives to conventional diagnostic methods.

This study focuses on leveraging interpretable machine learning algorithms to develop a predictive model for heart disease risk. We utilize a Decision Tree classifier for its straightforward visualization of decision rules, and we compare its performance with that of a Gaussian Naïve Bayes classifier, which provides a probabilistic perspective. Using the well-established Cleveland dataset as our experimental foundation, our methodology encompasses rigorous data cleaning, normalization, feature engineering, and comprehensive model evaluation. Our goal is to empower clinicians and researchers to make data-driven decisions through a practical and transparent framework.

II. Literature Review

A strong literature review is essential to understand the current state of research and identify gaps that our work addresses. In heart disease prediction, early studies compared various classifiers—such as Decision Trees, Naïve Bayes, k-Nearest Neighbors, and Random Forests—to determine their effectiveness on datasets like Cleveland.

Overview of Existing Research: Shah et al. [1] performed comparative analyses of several classifiers and emphasized that robust data preprocessing and effective feature selection are critical for achieving high predictive performance. Similarly, research by Dwivedi [2] and Chauhan et al. [3] confirmed that conventional models, when properly optimized, yield reliable results on the Cleveland dataset. While these studies demonstrate that even simpler models can perform well, many focus solely on accuracy without addressing the need for interpretability.

Critical Evaluation: Advanced ensemble and hybrid models, such as Random Forests and gradient boosting, have achieved impressive accuracy rates. However, their complex "black-box" nature often limits clinical applicability because they do not provide clear insights into the decision-making process. In contrast, simpler models like Decision Trees and Gaussian Naïve Bayes offer greater transparency. Although Gaussian Naïve Bayes typically exhibits higher accuracy, its assumption of conditional independence among features may oversimplify complex clinical relationships.

Identification of the Knowledge Gap: The literature reveals a crucial gap: the trade-off between achieving high accuracy and maintaining interpretability. Many existing studies either favor complex models that sacrifice explainability or focus on simple models without fully optimizing predictive performance. Our study addresses this gap by directly comparing a Decision Tree classifier (which offers clear, visual interpretability) with a Gaussian Naïve Bayes model (which provides a straightforward probabilistic framework). This dual-model approach aims to balance accuracy with the interpretability required for clinical decision support.

Synthesis: By integrating insights from previous studies and critically evaluating the strengths and weaknesses of various methods, our work demonstrates that achieving a balance between predictive performance and model transparency is both necessary and feasible. Our revised literature review emphasizes these points, setting the stage for our contribution to the field of heart disease prediction.

III. Methodology and Experimentation

A. Dataset and Preprocessing

We employ the Cleveland heart disease dataset obtained from the UCI repository via Kaggle, which comprises 303 records with 13 clinical predictors and one target variable indicating heart disease risk. Preprocessing steps included:

- **Missing Value Handling:** Imputation using the median was performed to address missing numerical values.
- **Normalization:** Min-Max scaling was applied to transform all features into the [0, 1] range.
- **Train-Test Split:** Stratified sampling was used to divide the dataset into 80% training and 20% testing sets, preserving the class distribution.

- **Feature Engineering:** A new composite feature, age_chol_ratio, was created by computing the ratio of age to cholesterol, capturing the combined influence of these two variables.

B. Model Implementation

Two machine learning models were developed using Python's scikit-learn library:

1. **Decision Tree Classifier:** The Decision Tree algorithm recursively partitions the dataset by selecting the feature that maximizes information gain. The entropy *H(S)*of a dataset *S* is defined as:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

where p_i is the probability of class i in *S*. For a feature *A* with values v, the information gain *IG(S,A)* is computed as:

$$IG(S,A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

Here, S_v represents the subset of *S* where feature *A* equals *v*. We tuned hyperparameters such as max_depth and min_samples_leaf to mitigate overfitting and ensure model generalizability.

2. **Gaussian Naïve Bayes Classifier:** Naïve Bayes uses Bayes' Theorem to compute the probability of a class *C* given a feature vector $X=(x_1,x_2,\ldots,x_n)$:

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)}$$

Under the assumption of conditional independence, the likelihood *P(X/C)* is expressed as:

$$P(X \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$$

For continuous features, assuming a Gaussian distribution yields:

$$P(x_i \mid C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(x_i - \mu_C)^2}{2\sigma_C^2}\right)$$

where μ_c and $\sigma^2{}_C$ are the mean and variance of feature x_i for class *C*.

C. Visualization and Evaluation

We assessed model performance using accuracy, precision, recall, and F1-score. Additionally, Principal Component Analysis (PCA) was applied to reduce the dataset to two dimensions, facilitating visual analysis:

- **Figure 1:** A Decision Tree diagram (generated using plot_tree) illustrates the hierarchical decision paths based on critical features.
- **Figure 2:** A PCA projection provides a two-dimensional visualization of the data distribution, highlighting both clustering and overlap among risk groups.
- **Figure 3 :** A flowchart summarizing the entire ML pipeline, from data import to visualization, reinforces the reproducibility of our methodology.

IV. Results and Discussion

A. Performance Comparison

On the testing set, the models yielded the following metrics:

- **Decision Tree Classifier:**

- ○ **Accuracy:** 77.05%
- ○ **Precision:** 0.73 (Class 0) and 0.82 (Class 1)
- ○ **Recall:** 0.83 (Class 0) and 0.72 (Class 1)
- ○ **F1-Score:** 0.77
- **Gaussian Naïve Bayes Classifier:**
- ○ **Accuracy:** 83.61%
- ○ **Precision:** 0.81 (Class 0) and 0.87 (Class 1)
- ○ **Recall:** 0.86 (Class 0) and 0.81 (Class 1)
- ○ **F1-Score:** 0.84

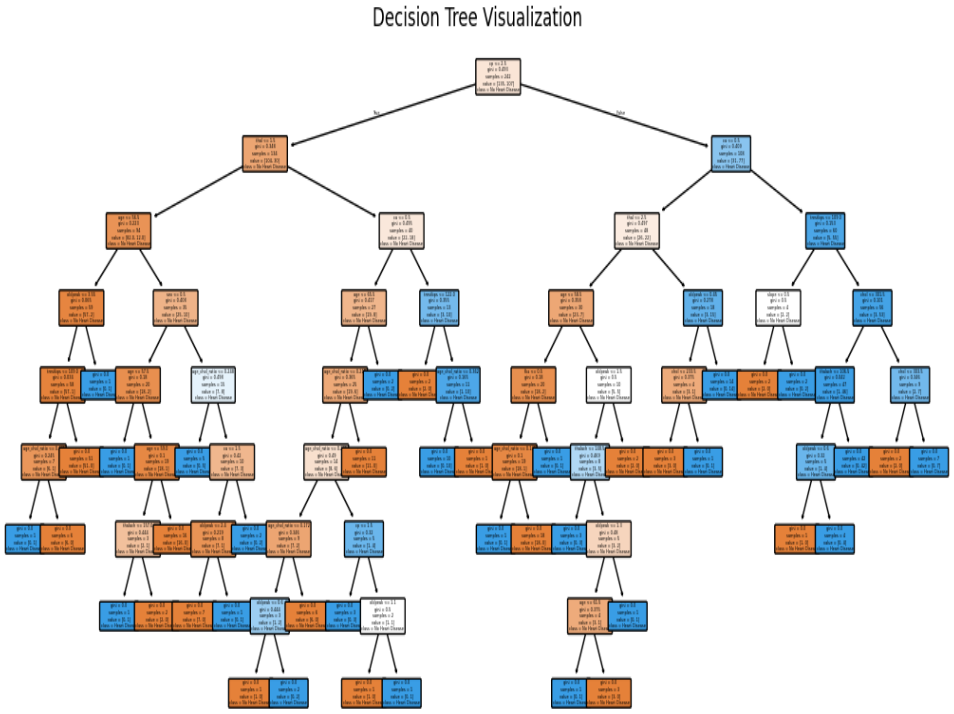
Table 1. Performance Comparison of Decision Tree and Naïve Bayes

Algorithm	Accuracy (%)	Precision (Class 0/1)	Recall (Class 0/1)	F1-Score
Decision Tree	77.05	73 / 82	83 / 72	0.77
Gaussian Naïve Bayes	83.61	81 / 87	86 / 81	0.84

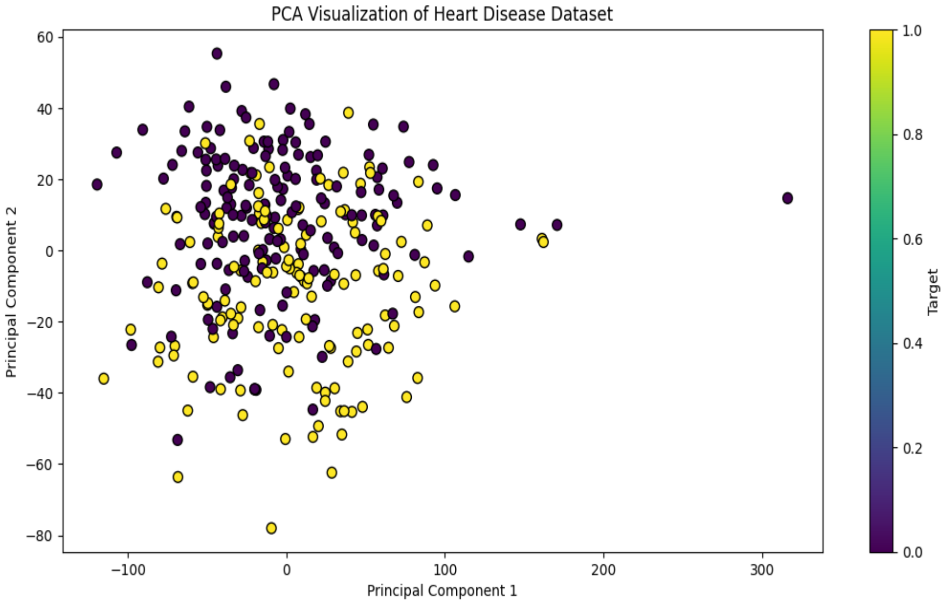
Note: Precision and recall values are reported for Class 0 and Class 1, respectively.

B. Visualization

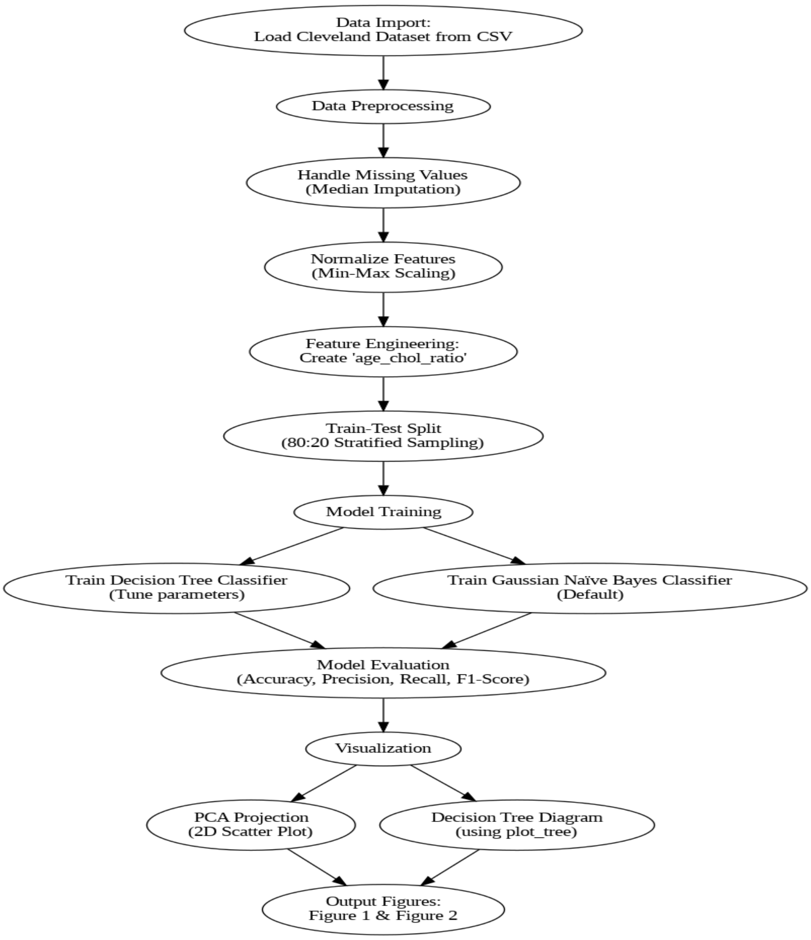
- **Figure 1: Decision Tree Diagram:** The diagram delineates the hierarchical structure of the Decision Tree, showcasing key splits based on features such as age, blood pressure, and cholesterol. This visual insight aids clinicians in understanding how risk factors converge to generate predictions.



- **Figure 2: PCA Projection:** The PCA projection offers a two-dimensional view of the dataset, revealing clusters of high-risk and low-risk patients, while also exposing overlapping regions that highlight the complexity of clinical data.



- **Figure 3:** A workflow diagram summarizing the complete ML pipeline can further enhance reproducibility and clarity.



C. Discussion

The Gaussian Naïve Bayes classifier demonstrates superior accuracy; however, the Decision Tree's transparent decision paths provide critical interpretability that is highly valued in clinical contexts. The overlapping clusters in the PCA projection indicate that some risk factors are common across patients, which underlines the inherent complexity of heart disease diagnosis. The trade-off between accuracy and interpretability is pivotal in healthcare: while higher accuracy models are desirable, a model that offers clear reasoning behind its predictions is crucial for clinical decision-making. This balance is the core strength of our approach, as it empowers clinicians to understand and trust the predictions made by the ML model.

V. Conclusion

This research presents a robust machine learning framework for predicting heart disease risk using the Cleveland dataset. By comparing a Decision Tree classifier with a Gaussian Naïve Bayes model, we have demonstrated that although the Naïve Bayes classifier achieves higher accuracy (83.61%), the interpretability of the Decision Tree is essential for clinical application. Our methodology—incorporating meticulous data preprocessing, thoughtful feature engineering, and comprehensive model evaluation—provides a practical, reproducible solution for healthcare analytics. Future research will explore hybrid models and integrate additional clinical features, along with external validation, to further enhance the model's predictive power and generalizability.

References

1. M. D. Seckeler and T. R. Hoke, "The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease," *Clin. Epidemiol.*, vol. 3, p. 67, 2011.
2. D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, p. 345, 2020.
3. A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, vol. 29, no. 10, pp. 685–693, 2018.
4. R. Chauhan, P. Bajaj, K. Choudhary, and Y. Gigras, "Framework to predict health diseases using attribute selection mechanism," in *Proc. 2nd Int. Conf. Comput. Sustain. Global Dev. (INDIACom)*, 2015, pp. 1880–1884.
5. K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart diseases detection using Naïve Bayes algorithm," *Int. J. Innov. Sci. Eng. Technol.*, vol. 2, no. 9, pp. 441–444, 2015.