The impact of data preprocessing and feature selection on machine learning-based cardiovascular disease prediction

Kyana Marckx

January 31, 2026

1 Introduction

Cardiovascular diseases (also known as heart diseases/failures) remain one of the leading causes of mortality worldwide, accounting for a substantial proportion of premature deaths and long-term morbidity. Early detection of heart disease is therefore of critical importance, as timely intervention can significantly reduce the risk of severe complications and improve patient outcomes. With the increasing availability of clinical and demographic data, machine learning has emerged as a promising tool to support medical decision-making by identifying complex patterns that may not be evident through traditional statistical approaches.

In recent years, numerous studies have explored the application of machine learning techniques for heart disease prediction. Comparative analyses show that supervised learning models such as Random Forests, gradient boosting methods and neural networks can achieve strong predictive performance on publicly available datasets, often outperforming traditional statistical models [6][7]. These studies primarily focus on maximizing performance metrics such as accuracy or ROC-AUC through model comparison and hyperparameter optimization. Based on these findings, this project will only focus on training a Random Forest and an XGBoost since those dominate the research outcomes.

However, systematic reviews of machine learning applications in heart disease prediction indicate that many studies place limited emphasis on data preprocessing and feature selection, despite their significant influence on model performance and interpretability [2]. In several cases, preprocessing steps are either insufficiently described or treated as secondary to algorithmic choice, which can hinder reproducibility and obscure the true sources of performance gains.

The importance of preprocessing is particularly evident in clinical datasets, where raw measurements may contain implausible or inconsistent values. Medical literature highlights that extreme values for physiological indicators such as blood pressure and cholesterol levels may reflect measurement errors rather than true clinical states. For instance, abnormally low blood pressure values or zero cholesterol measurements are physiologically unlikely and should be handled carefully using domain knowledge during data cleaning [1][3][4]. Ignoring such issues can negatively impact both model reliability and clinical relevance.

Recent work has also emphasized the importance of interpretability and transparency in medical machine learning applications, proposing explainable AI approaches alongside predictive modeling to support clinical trust and decision-making [5]. These considerations further reinforce the need for a well-documented and methodologically sound data science pipeline.

The objective of this project is therefore not to outperform existing approaches through extensive algorithmic experimentation, but to systematically investigate how preprocessing strategies and feature selection choices affect machine learning performance in heart disease prediction. By evaluating multiple feature subsets using a consistent modeling framework and cross-validation strategy, this project aims to provide clearer insight into the role of data preparation within supervised learning. In doing so, the project aligns closely with the objectives of the Machine Learning course, emphasizing methodological rigor, interpretability and a structured data science approach over purely performance-driven optimization.

2 Goal

The main objective of this project is to investigate how different data preprocessing and feature selection strategies influence the performance of a machine learning model for heart disease prediction. Rather than comparing a large number of algorithms, this project focuses on understanding the impact of data-centric decisions within a supervised learning pipeline.

2.1 Main research question

How do different preprocessing and feature selection choices affect the predictive performance of a machine learning model for heart disease prediction?

2.2 Sub-research objectives

To answer the main research question, the project is structured around the following smaller objectives:

- To analyze and preprocess a publicly available heart disease dataset by handling missing values, outliers and categorical variables using domain-informed decisions.
- To define and evaluate multiple feature subsets base on clinical relevance and exploratory data analysis.
- To train maximum two, literature-supported machine learning models using a consistent pipeline and cross-validation strategy. *(Note: Random Forest and XGBoost were chosen as the two models)*
- To evaluate the resulting models using appropriate classification metrics and to compare their performance across different preprocessing configurations.

The remainder of this report is structured as follows. First, the dataset and preprocessing steps are described and analyzed. Next, the methodological approach and implementation choices are discussed. This is followed by an evaluation of the trained models and an interpretation of the results. Finally, conclusions are drawn and limitations and directions for future research are outlined.

3 Data analysis

3.1 Dataset description

The dataset used in this project is the *Heart Failure Prediction Dataset*[1], a publicly available clinical dataset containing patient information relevant to cardiovascular health. The dataset consists of (indirectly) demographic, clinical and diagnostic features commonly associated with heart disease risk. The target variable indicates the presence or absence of heart disease, making this a binary classification problem.

The dataset includes both numerical features such as age, resting blood pressure, cholesterol level and maximum heart rate, and categorical features such as sex, chest pain type, resting ECG results, exercise-induced angina and ST slope. This mixture of feature types requires careful preprocessing to ensure compatibility with machine learning algorithms.

3.2 Data preprocessing and cleaning

Prior to modeling, the dataset was thoroughly inspected for data quality issues. No duplicate records were identified; however, several features contained implausible or invalid values. In particular, zero values for cholesterol levels and extremely low values for resting blood pressure were

[1] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

identified as physiologically unlikely and treated as invalid measurements. These values were handled through imputation strategies informed by both statistical reasoning and medical domain knowledge.

Missing or invalid numerical values were imputed using robust statistical measures, while categorical variables were encoded using one-hot encoding. Numerical features were not scaled because Random Forest and XGBoost are both tree-based models and therefore do not rely on distance-based metrics, making scaling unnecessary. All preprocessing steps were implemented using reproducible pipelines to avoid data leakage and to ensure consistency across experiments.

3.3 Outlier analysis

Outlier analysis was conducted on numerical features to identify extreme values that could disproportionately influence model training. Rather than removing outliers indiscriminately, each feature was evaluated in the context of clinical plausibility. Values deemed physiologically possible were retained, while clearly erroneous measurements were treated as missing and imputed accordingly. This approach preserves meaningful clinical variability while reducing noise introduced by data errors.

3.4 Exploratory Data Analysis (EDA) and visualization

Exploratory Data Analysis was performed using a combination of univariate and multivariate visualizations, including kernel density plots, boxplots, pairplots and correlation matrices. These analyses provided insight into feature distributions, relationships between variables and potential associations with the target variable.

Visual inspection revealed notable differences in several features between patients with and without heart disease, supporting their relevance for predictive modeling. Correlation analysis further highlighted relationships among numerical variables, informing subsequent feature selection decisions. Selected plots and detailed observations are provided separately to support interpretability and reproducibility.

3.5 Feature subsets

Based on the EDA and domain considerations, multiple feature subsets were defined to evaluate the influence of feature selection on model performance. These subsets include combinations of clinically relevant features and subsets derived from exploratory insights. Evaluating these subsets enables a systematic comparison of how feature inclusion affects predictive performance, while maintaining a consistent modeling framework.

4 Methodology and implementation

5 Evaluation and results

6 Conclusions and discussion

7 Tables and figures

7.1 Tables

7.2 Figures

References

[1] A. H. Association. Low blood pressure - when blood pressure is too low.

[2] T. Banerjee and I. Paçal. A systematic review of machine learning in heart disease prediction. *Turkish Journal of Biology = Turk Biyoloji Dergisi*, 49(5):600–634, 2025.

[3] C. Clinic. What should my cholesterol levels be?

[4] M. Clinic. Isolated systolic hypertension: A health concern?

[5] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif. A proposed technique for predicting heart disease using machine learning algorithms and an explainable ai method. *Scientific Reports*, 14(1):23277, Oct. 2024.

[6] A. R. Ilyas, S. Javaid, and I. L. Kharisma. Heart disease prediction using ml. In *The 7th International Global Conference Series on ICT Integration in Technical Education and Smart Society*, page 124. MDPI, Oct. 2025.

[7] M. D. Teja and G. M. Rayalu. Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance. *BMC Cardiovascular Disorders*, 25(1):212, Mar. 2025.