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

Kyana Marckx - 852793982

January 31, 2026

Keywords – preprocessing, feature selection, supervised machine learning, cardiovascular diseases

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 two 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 based on clinical relevance and exploratory data analysis.
- To train a maximum of two literature-supported machine learning models using a consistent pipeline and cross-validation strategy. *(Update/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 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 (median values), while categorical variables were encoded using one-hot encoding (and an imputation with a *most frequent* strategy, but in this dataset there were no missing values so it's an extra step for future usage). 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.

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

3.4 Exploratory Data Analysis (EDA) and visualization

Exploratory Data Analysis was performed using a combination of univariate, bivariate 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

4.1 Research methodology

This project follows a structured data science methodology aligned with standard machine learning workflows. The approach consists of sequential stages including Exploratory Data Analysis, data preprocessing, feature selection, model training and evaluation. Particular emphasis is placed on isolating the impact of preprocessing and feature selection choices while keeping the modeling approach consistent across experiments.

Rather than performing extensive model benchmarking, a limited number of well-established supervised learning algorithms were selected based on prior literature. This design choice allows for a controlled analysis in which variations in model performance can be attributed primarily to differences in data preparation rather than algorithmic complexity.

4.2 Design and pipeline architecture

To ensure reproducibility and prevent data leakage, preprocessing and modeling steps were combined into unified machine learning pipelines using the `scikit-learn` framework. Numerical and categorical features were processed separately using a `ColumnTransformer`, enabling tailored preprocessing for each feature type.

Numerical preprocessing included imputation of missing or invalid values. Categorical features were transformed using one-hot encoding to allow their inclusion in tree-based and ensemble models. Multiple feature subsets were incorporated into the pipeline design, allowing systematic evaluation of different feature combinations under identical preprocessing and modeling conditions.

All preprocessing pipelines were saved and reused during model training to ensure consistency across experiments.

4.3 Model selection and training

Based on findings from existing literature, tree-based ensemble methods were selected due to their strong performance in heart disease prediction tasks and their ability to handle nonlinear relationships and mixed feature types. In particular, Random Forest and Gradient Boosting-based models were evaluated, with Random Forest emerging as the best-performing model across most feature subsets.

Model training was performed using cross-validation to obtain robust performance estimates and reduce the risk of overfitting. Hyperparameter tuning was conducted using grid search within the cross-validation framework. A stratified splitting strategy was employed to preserve class distributions during training and testing.

4.4 Implementation details

The entire solution was implemented in Python using Jupyter Notebooks, available on GitHub[2]. Core libraries included `pandas` and `numpy` for data-handling, `scikit-learn` for preprocessing, modeling and evaluation, and additional libraries for visualization. Model artifacts, including trained pipelines, were stored using `joblib` to support reproducibility and separation between preprocessing, modeling and evaluation stages.

By structuring the implementation around reusable pipelines and clearly defined experimental configurations, this project ensures transparency, reproducibility and methodological rigor, which are essential for machine learning applications in healthcare contexts.

5 Evaluation and results

5.1 Evaluation methodology

Model performance was evaluated using a held-out test set that was not used during training or hyperparameter tuning. During the modeling phase, cross-validation was applied to obtain robust performance estimates. Given the binary classification setting and the clinical relevance of correctly identifying heart disease cases, multiple evaluation metrics were considered.

The primary evaluation metric in this project is the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), as it provides a threshold-independent measure of model discrimination. In addition, precision, recall, F1-score, accuracy, confusion matrices and precision-recall curves were used to obtain a comprehensive view of model behavior.

5.2 Results

Across all evaluated preprocessing configurations and feature subsets (see Table 1 for the columns inside each feature subset and Table 2 for the evaluation metrics of all trained models), the Random Forest model achieved the strongest overall performance. Used parameters after finetuning with 10-fold cross-validation are visible in Table 3. The best-performing configuration obtained a ROC-AUC of approximately 0.88 on the test set, as shown in Figure 1, indicating strong discriminative ability between patients with and without heart disease.

The ROC curve in Figure 1 lies well above the diagonal line representing random guessing, demonstrating that the model achieves high true positive rates at relatively low false positive rates. This confirms that the model is effective at separating positive and negative cases across a wide range of classification thresholds.

Further insights into model performance is provided by the precision-recall curve in Figure 2. The curve shows that precision remains high over a broad range of recall values, indicating that the model is able to identify most patients with heart disease while keeping the number of false positive predictions limited. Precision decreases only when recall approaches its maximum, reflecting the expected trade-off when attempting to capture nearly all positive cases.

The confusion matrix in Figure 3 summarizes the classification outcomes on the test set. The majority of both positive and negative cases are correctly classified, with a relatively small number of false negatives and false positives. This observation is consistent with the quantitative metrics reported in Table 4, which shows high precision, recall and F1-scores for both classes, as well as an overall accuracy of 0.89.

5.3 Interpretation

The results illustrated in Figure 1, Figure 2, Figure 3 and Table 4 demonstrate that preprocessing strategies and feature selection choices have a substantial impact on model performance, even when using a fixed learning algorithm. Feature subsets informed by Exploratory Data Analysis and clinical relevance consistently led to improved performance compared to less selective feature combinations.

From a clinical perspective, the strong recall for the positive class shown in Table 4 is particularly important, as it indicates that most heart disease cases are correctly identified. While some false positives remain, as visualized in Figure 3, their number remains limited, suggesting that the model may be suitable for use as a decision-support or screening tool rather than a definitive diagnostic system.

Overall, the evaluation results confirm that careful data preprocessing and feature selection contribute significantly to robust and reliable machine learning performance in heart disease prediction, reinforcing the importance of data-centric modeling approaches in healthcare applications.

[2] https://github.com/kyanamarckx/cardiovascular-diseases

6 Conclusions and discussion

This project investigated the impact of data preprocessing and feature selection on machine learning performance for heart disease prediction. Rather than focusing on extensive algorithm comparison, the study adopted a data-centric approach to assess how preprocessing decisions influence predictive outcomes within a supervised learning framework.

The results demonstrate that preprocessing strategies and feature selection play a critical role in model performance. Carefully handling invalid or implausible values, encoding categorical variables and selecting clinically relevant feature subsets consistently led to improved and more stable performance. Using a fixed modeling approach allowed differences in results to be attributed primarily to data preparation choices rather than algorithmic variation.

The Random Forest model achieved strong predictive performance, with high ROC-AUC, precision and recall values across the evaluated configurations. In particular, the model's ability to correctly identify patients with heart disease while maintaining a limited number of false positives highlights its suitability as a decision-support or screening tool. These findings align with existing literature, which reports strong performance of ensemble-based models in heart disease prediction, while extending prior work by explicitly demonstrating the influence of preprocessing choices.

Despite these promising results, several limitations should be acknowledged. The dataset used is relatively small and originates from a single publicly available source, which may limit generalizability to broader or more diverse populations. Additionally, while domain knowledge was incorporated during data cleaning, some preprocessing decisions remain inherently subjective. Finally, this project focused on predictive performance rather than model explainability, which is an important consideration in clinical applications.

Future research could address these limitations by validating the findings on larger and more diverse datasets, incorporating explainable AI techniques to enhance interpretability and exploring the interaction between preprocessing strategies and other model families. Furthermore, collaboration with clinical experts could further refine preprocessing decisions and strengthen the practical relevance of the model.

In conclusion, this project highlights that thoughtful preprocessing and feature selection are essential components of effective machine learning pipelines in healthcare. By emphasizing methodological rigor and reproducibility, the study reinforces the importance of data-centric approaches in applied machine learning, in line with the objectives of the Machine Learning course.

7 Tables and figures

7.1 Tables

Name	Columns
all	Numerical & categorical features
numerical	Numerical features
categorical	Categorical features
FS-1	MaxHR, Oldpeak, ST_Slope, ChestPainType
FS-2	FS-1 + RestingECG & Age
FS-3	FS-2 + Cholesterol & ExerciseAngina
paper	Sex, ChestPainType, RestingBP, FastingBS, RestingECG, ExerciseAngina, ST_Slope

Table 1: Columns in feature subsets

Model	Subset	Folds	Accuracy	ROC-AUC	F1-score
Random Forest	all	10	0.886957	0.937619	0.899225
XGBoost	all	10	0.891304	0.937161	0.901961
Random Forest	paper	10	0.873913	0.931848	0.888031
XGBoost	paper	10	0.860870	0.924509	0.878788
Random Forest	categorical	10	0.860870	0.923744	0.874016
XGBoost	FS-3	10	0.869565	0.921948	0.880952
XGBoost	categorical	10	0.860870	0.921222	0.875000
Random Forest	FS-3	5	0.860870	0.920725	0.874016
XGBoost	FS-1	10	0.852174	0.913386	0.865079
XGBoost	FS-2	10	0.852174	0.913309	0.865079
Random Forest	FS-1	5	0.821739	0.906850	0.837945
Random Forest	FS-2	10	0.817391	0.906506	0.832000
XGBoost	numerical	5	0.804348	0.868359	0.820717
Random Forest	numerical	5	0.782609	0.867136	0.796748

Table 2: Evaluation metrics for all trained models

Subset	max_depth	min_samples_leaf	n_estimators
all	None	5	300

Table 3: Finetuned parameters for Random Forest

Heart Disease	Precision	Recall	F1-score
0 (no)	0.89	0.85	0.87
1 (yes)	0.89	0.91	0.90

Table 4: Classification report of Random Forest

7.2 Figures

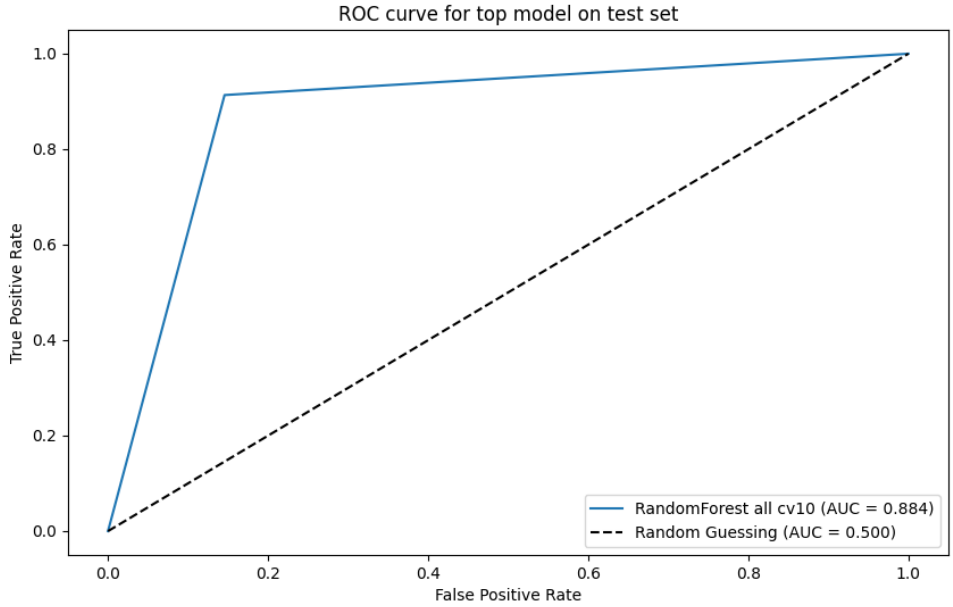

Figure 1: ROC curve for top model on test set

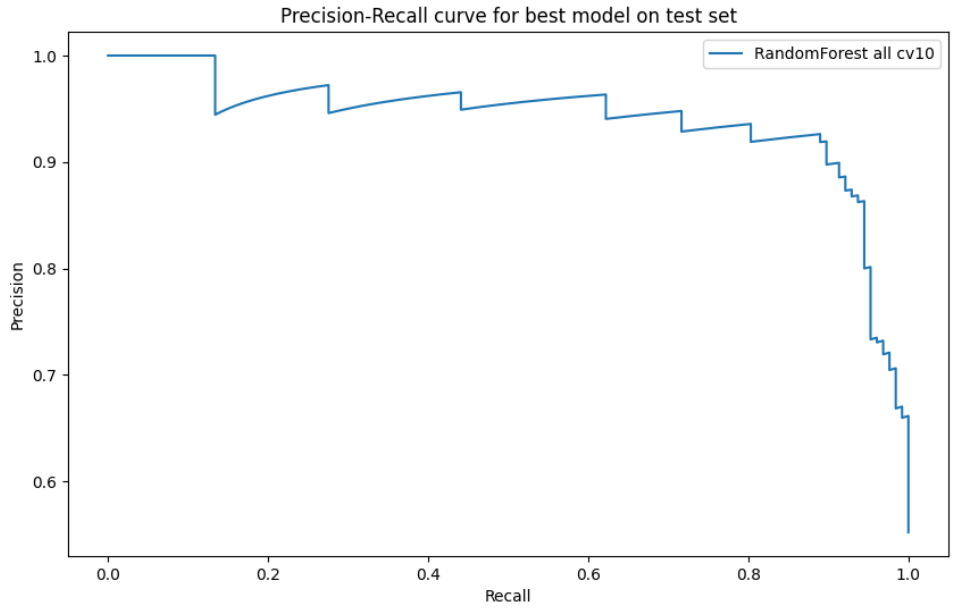

Figure 2: Precision-Recall curve for best model on test set

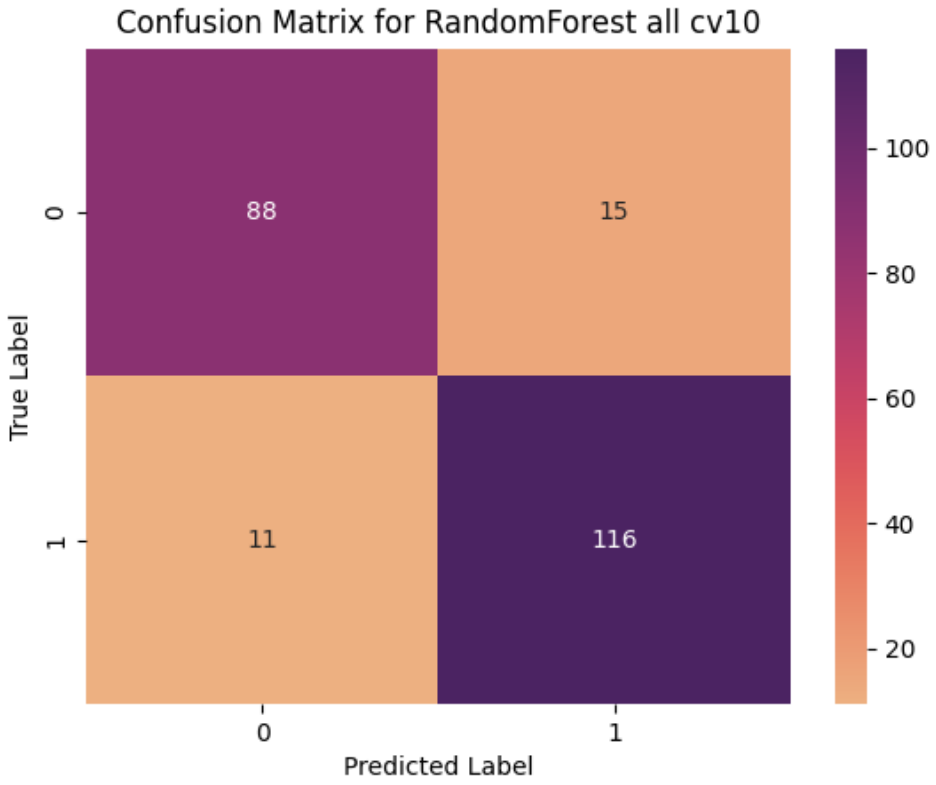

Figure 3: Confusion matrix for best model on test set

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.