



**ÉCOLE SUPÉRIEURE D'INGÉNIEURS
LÉONARD-DE-VINCI**

Heart Disease Prediction Using Machine Learning

**Nassim Loudiyi
Paul-Adrien Lu-Yen-Tung**

ESILV – M1 Data & Artificial Intelligence - DIA4
Academic Year 2025–2026

Supervisor: Ali Mokh

December 2025

ABSTRACT

Cardiovascular diseases represent a major global health challenge, making early detection essential for prevention and clinical decision support. This project explores supervised machine learning approaches to predict heart disease based on structured clinical and physiological variables. The methodology follows a complete end-to-end pipeline including data quality assessment, preprocessing, baseline modeling, hyperparameter tuning, ensemble learning, and model comparison.

Tree-based ensemble algorithms — particularly CatBoost — achieved the highest predictive performance, with strong ROC-AUC scores and robust generalization across evaluation metrics. SHAP-based interpretability techniques were then employed to quantify individual and global feature contributions, providing medically meaningful insights into the primary determinants of cardiovascular risk.

Overall, the results demonstrate the effectiveness of gradient-boosting models for reliable and interpretable heart disease prediction.

Contents

1	Introduction	2
	Introduction	2
2	Dataset and Problem Definition	2
2.1	Clinical Context	2
2.2	Prediction Task	3
2.3	Dataset Description and Features	4
2.4	Summary Statistics	5
3	Step 1: Data Quality, EDA and Preprocessing	6
3.1	Data Quality Checks	6
3.1.1	Missing Values and Duplicate Rows	6
3.1.2	Outliers in Continuous Variables	7
3.1.3	Class Imbalance	8
3.2	Exploratory Data Analysis	8
3.2.1	Numerical Features	9
3.2.2	Pairwise Relationships Between Numerical Features	10
3.2.3	Categorical Features	12
3.2.4	Correlation Analysis	13
3.3	Principal Component Analysis (PCA)	14
3.4	Train–Test Split and Preprocessing Pipeline	15
3.5	Baseline Models	16
4	Step 2: Modeling and Hyperparameter Tuning	20
4.1	Default Model Benchmark	21
4.2	Advanced Model — CatBoost	23
4.3	Tuned Models	24
4.4	Learning and Validation Curves	25
4.4.1	Learning Curves	25
4.4.2	Validation Curves	27
4.5	Ensemble Models	28
4.6	Final Model Comparison	29
5	Step 3: Explainability with SHAP	31
5.1	Global Interpretability	32
5.2	Local Interpretability	33
6	Discussion	35
6.1	Results Interpretation	35
6.2	Limitations	35
6.3	Ethical Considerations	36
7	Conclusion and Future Work	36
	References	38

1 Introduction

Cardiovascular diseases remain one of the leading causes of mortality worldwide and represent a major public health challenge for both prevention and early clinical intervention. Early identification of individuals at high risk of developing heart disease is essential for improving treatment outcomes, reducing hospitalization costs, and supporting clinical decision-making. In recent years, the rise of data-driven approaches has provided new opportunities to extract predictive patterns from structured clinical data and assist clinicians in diagnosing cardiac conditions.

Machine learning techniques have demonstrated strong potential in the medical domain by enabling the analysis of complex, multidimensional data and capturing non-linear interactions between clinical variables. When properly designed and validated, such models can improve risk stratification, highlight relevant prognostic factors, and complement standard diagnostic procedures. However, developing reliable medical prediction systems requires rigorous methodology, including data quality assessment, appropriate preprocessing, careful model selection, robust evaluation, and transparent interpretability.

This project aims to design, evaluate, and interpret a complete supervised machine learning pipeline for heart disease prediction using a structured clinical dataset. The study follows a standard three-stage methodology. First, we conduct data quality checks, exploratory data analysis, and preprocessing, while establishing strong baseline models. Second, we benchmark several classification algorithms, perform hyperparameter tuning, analyze learning and validation curves, and evaluate ensemble strategies to identify the most effective predictive model. Finally, we apply SHAP-based explainability techniques to understand global and local feature contributions, ensuring interpretability and clinical relevance.

By combining rigorous data exploration, advanced predictive modeling, and state-of-the-art explainability methods, this project provides a comprehensive evaluation of automated heart disease prediction. The results highlight both the performance of modern ensemble methods and the importance of transparency when deploying machine learning systems in medical contexts.

2 Dataset and Problem Definition

2.1 Clinical Context

Cardiovascular diseases are among the primary causes of morbidity and mortality worldwide, accounting for millions of deaths each year. Heart disease, in particu-

lar, encompasses a wide range of conditions that affect the heart's structure and function, including coronary artery disease, arrhythmias, and heart failure. Early diagnosis is crucial, as timely intervention significantly improves patient prognosis and reduces long-term complications.

Clinical assessment traditionally relies on a combination of medical history, physical examination, laboratory tests, and diagnostic tools such as electrocardiograms and imaging techniques. While these methods are effective, they can be time-consuming, resource-intensive, and subject to inter-clinician variability. Furthermore, many risk factors interact in complex, non-linear ways that may be difficult to identify through manual analysis alone.

The increasing availability of structured clinical datasets has opened new opportunities for leveraging machine learning to support cardiovascular risk assessment. By analyzing patterns within physiological measurements and symptoms, machine learning models can help identify high-risk patients earlier and complement the decision-making process of healthcare professionals.

2.2 Prediction Task

The goal of this project is to build a supervised machine learning model capable of predicting the presence of heart disease based on a set of clinical and physiological attributes. The problem is formulated as a **binary classification task**, where each patient is assigned a label indicating whether heart disease is present ($y = 1$) or absent ($y = 0$).

Given a feature vector $X = (x_1, x_2, \dots, x_n)$ describing a patient's medical characteristics, the objective is to learn a function:

$$f : X \rightarrow \{0, 1\}$$

that maps the input features to the correct diagnostic outcome. The dataset contains both continuous and categorical variables, requiring appropriate preprocessing steps such as scaling, encoding, and handling of potential outliers.

The predictive model aims to detect clinically relevant patterns associated with cardiovascular risk, complementing traditional diagnostic procedures. Performance will be evaluated using metrics suited for medical classification tasks, including accuracy, precision, recall, F1-score, and ROC–AUC, ensuring that the selected model is both reliable and clinically meaningful.

2.3 Dataset Description and Features

The dataset used in this study contains structured clinical information collected from patients undergoing cardiovascular examinations. It comprises a mix of demographic variables, medical measurements, electrocardiographic indicators, and exercise-induced attributes. Each row corresponds to one patient, and each column represents a specific clinical feature or diagnostic outcome.

The dataset includes a total of 13 predictive features, along with a binary target variable indicating the presence of heart disease. These features capture several aspects of cardiovascular health, including blood pressure, cholesterol levels, heart rate response to exercise, chest pain type, and abnormalities detected through electrocardiography. The diversity of feature types requires appropriate preprocessing steps, such as numerical scaling and categorical encoding.

Table 1 summarizes the input features, their clinical meaning, and their data types.

Feature	Description	Type
patientid	Unique patient identification number	Numeric
age	Patient age in years	Continuous
gender	Sex (1 = male, 0 = female)	Binary
chestpain	Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)	Categorical
restingBP	Resting blood pressure (94–200 mmHg)	Continuous
serumcholesterol	Serum cholesterol level (126–564 mg/dl)	Continuous
fastingbloodsugar	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)	Binary
restingelectro	Resting ECG results (0 = normal, 1 = ST-T abnormality, 2 = LV hypertrophy)	Categorical
maxheartrate	Maximum heart rate achieved (71–202 bpm)	Continuous
exerciseangia	Exercise-induced angina (1 = yes, 0 = no)	Binary
oldpeak	ST depression induced by exercise	Continuous
slope	Slope of the ST segment (1 = upsloping, 2 = flat, 3 = downsloping)	Categorical
noofmajorvessels	Number of major vessels visualized (0–3)	Numeric
target	Heart disease classification (1 = disease, 0 = no disease)	Binary

Table 1: Clinical features included in the dataset, fully aligned with the official dataset description.

This combination of heterogeneous clinical features makes the dataset well suited for machine learning applications, as it captures multiple dimensions of cardiovascular function and patient health. The next sections provide a statistical overview and quality assessment of the dataset.

2.4 Summary Statistics

A preliminary statistical summary was computed for all numerical and categorical features to obtain an overview of the dataset's structure, central tendencies, and dispersion. This descriptive analysis helps identify potential anomalies, assess feature variability, and understand the clinical ranges represented in the population. These statistics serve as a foundation for subsequent quality checks, exploratory analysis, and preprocessing decisions.

Feature	Count	Mean	Std	Min	25%	50%	75%
age	1000	49.24	17.86	20	34.00	49.00	64.25
gender	1000	0.76	0.42	0	1.00	1.00	1.00
chestpain	1000	0.98	0.95	0	0.00	1.00	2.00
restingBP	1000	151.47	29.97	94	129.00	147.00	181.00
serumcholestrol	1000	311.45	132.44	0	236.00	318.00	404.00
fastingbloodsugar	1000	0.29	0.45	0	0.00	0.00	1.00
restingelectro	1000	0.78	0.77	0	0.00	1.00	1.00
maxheartrate	1000	145.47	29.41	71	120.00	146.00	175.00
exerciseangia	1000	0.49	0.50	0	0.00	0.00	1.00
oldpeak	1000	2.27	1.72	0	1.30	2.40	4.10
slope	1000	1.54	1.00	0	1.00	2.00	2.00
noofmajorvessels	1000	1.22	0.98	0	1.00	1.00	2.00
target	1000	0.58	0.49	0	0.00	1.00	1.00

Table 2: Descriptive summary statistics for numerical and categorical features.

The summary indicates substantial variability across several clinical variables. Resting blood pressure and cholesterol levels show wide ranges, reflecting heterogeneous cardiovascular profiles within the population. The *maxheartrate* distribution aligns with physiological expectations, while *oldpeak* values show typical patterns of ST depression associated with exercise-induced stress.

Categorical variables such as *gender*, *chestpain*, and *restingelectro* display diverse distributions consistent with real-world clinical cohorts. The target variable has a mean

3 Step 1: Data Quality, EDA and Preprocessing

This first stage of the project focuses on understanding the structure and integrity of the dataset before developing any predictive model. High-quality data is essential for building reliable machine learning solutions, especially in medical applications where noisy or inconsistent observations may lead to misleading conclusions.

Step 1 is structured into three main components. First, we perform a comprehensive **data quality assessment**, including checks for missing values, duplicate entries, outliers in continuous variables, and potential imbalance in the target distribution. Second, we conduct an **exploratory data analysis (EDA)** to better understand the statistical properties of the features, visualize their distributions, and inspect relationships between variables through histograms, countplots, and correlation matrices. Finally, we define the **preprocessing pipeline** required to prepare the data for modeling, including train-test splitting, encoding of categorical features, scaling of numerical variables, and a preliminary PCA exploration.

Together, these steps ensure that the dataset is clean, well understood, and properly formatted for the modeling phase that follows in Step 2.

3.1 Data Quality Checks

A preliminary quality assessment was performed to evaluate the integrity and suitability of the dataset before conducting exploratory analysis and model training. This section examines four essential aspects: missing values, duplicate rows, outliers in continuous variables, and class imbalance.

3.1.1 Missing Values and Duplicate Rows

The inspection of missing values shows that none of the clinical features contain incomplete entries. This is particularly advantageous in a medical context, as imputation can distort physiological information or introduce artificial patterns.

Similarly, the dataset was examined for duplicate patient profiles. No duplicated rows were detected, ensuring that each patient contributes uniquely to the dataset and that no observation is unintentionally overweighted.

```

    === Missing Values ===
patientid      0
age            0
gender         0
chestpain      0
restingBP      0
serumcholesterol 0
fastingbloodsugar 0
restingelectro 0
maxheartrate   0
exerciseangia  0
oldpeak        0
slope          0
noofmajorvessels 0
target         0
dtype: int64
Total missing values: 0

```

```

    === Duplicate Rows ===
Total duplicated rows: 0

```

(a) Missing Values

(b) Duplicate Rows

Figure 1: Data quality checks: missing values (a) and duplicate observations (b). Both analyses confirm that the dataset is complete and contains no repeated patient records.

3.1.2 Outliers in Continuous Variables

Outlier analysis was conducted on the continuous clinical variables (*age*, *restingBP*, *serumcholesterol*, *maxheartrate*, *oldpeak*). The boxplots in Figure 2 do not reveal any extreme outliers according to the standard $1.5 \times \text{IQR}$ rule. Although a few observations lie near the upper bounds of the distributions (e.g., cholesterol and ST depression), these values remain within clinically plausible ranges and are not flagged as statistical outliers.

This behaviour is expected, as the dataset originates from a curated clinical source in which extreme or erroneous measurements were likely removed during preprocessing. Consequently, all observations were retained to preserve the natural physiological variability present in the population.

Note. Outlier detection applies only to continuous variables, where quartiles and deviation thresholds are meaningful. Categorical features, representing discrete clinical states, do not admit outliers in the statistical sense and are instead analysed using frequency-based methods (countplots), presented later in the EDA section.

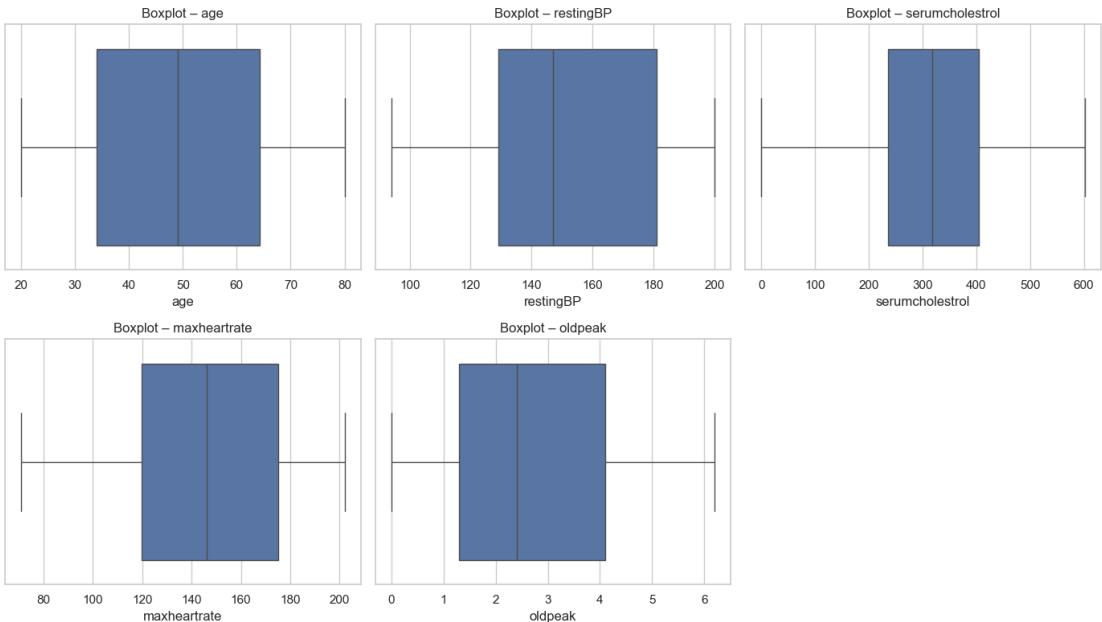


Figure 2: Boxplots of continuous clinical variables for outlier detection.

3.1.3 Class Imbalance

The distribution of the target variable was examined to identify potential class imbalance. As shown below, the dataset contains 58% positive cases (heart disease) and 42% negative cases, indicating a mild but manageable imbalance. Although not severe enough to require resampling, this difference justifies monitoring precision, recall, and ROC–AUC during model evaluation, as misclassification costs are asymmetric in a medical context.

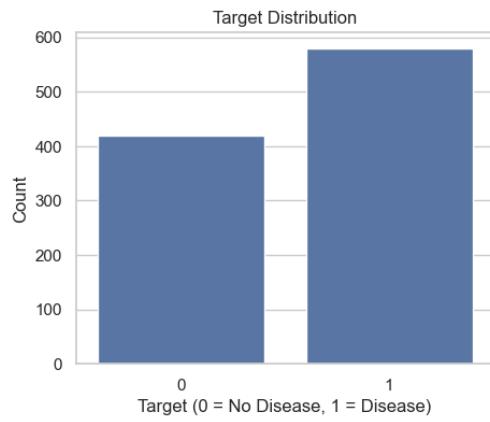
The slight predominance of class 1 indicates that the dataset contains more patients diagnosed with heart disease. While the imbalance is not severe enough to require resampling, the higher clinical cost of false negatives justifies careful monitoring of sensitivity-related metrics during model evaluation.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to better understand the distributional properties of the variables, detect potential irregularities, and study relationships between features and the target variable. Both numerical and categorical attributes were analyzed separately due to their different statistical nature.

==== Target Distribution ===		
	Count	Percent
target		
0	420	42.0
1	580	58.0

(a) Tabular summary of target distribution.



(b) Graphical representation of the target distribution.

Figure 3: Class imbalance analysis using both numerical output (a) and histogram plot (b).

3.2.1 Numerical Features

The numerical clinical variables (*age*, *restingBP*, *serumcholesterol*, *maxheartrate*, *oldpeak*) were examined using overlaid histograms with kernel density estimation (KDE). These plots allow simultaneous visualization of the distribution for patients with and without heart disease.

Several patterns emerge from these distributions. Patients diagnosed with heart disease tend to exhibit:

- slightly higher **resting blood pressure** and **cholesterol levels**;
- noticeably higher **maximum heart rate achieved**;
- higher **oldpeak** values, indicating more pronounced ST-segment depression.

These trends align with well-established cardiovascular risk factors and suggest that the numerical variables carry meaningful predictive information for classification.

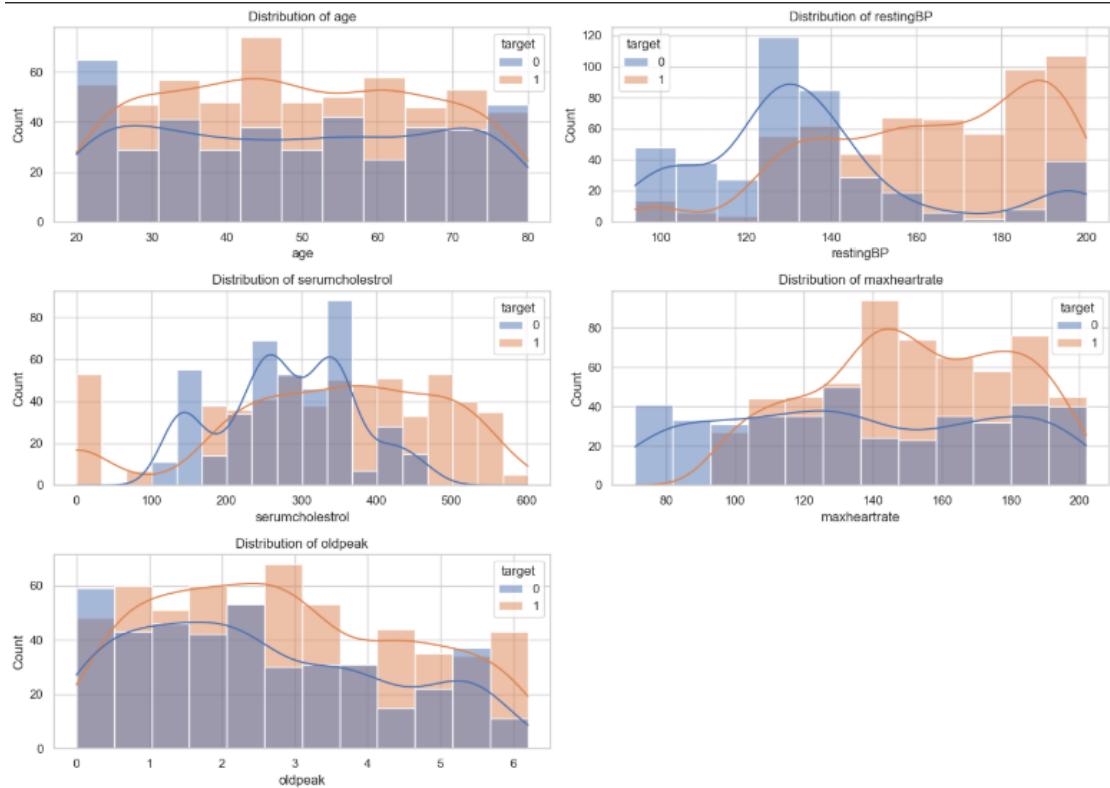


Figure 4: Distribution of numerical features for each class of the target variable.

3.2.2 Pairwise Relationships Between Numerical Features

To complement the univariate analysis, a pairplot was generated to examine pairwise relationships between numerical features and assess potential class separability in the raw feature space.

This visualization below highlights several important insights:

- The variables exhibit **no strong linear correlations**, confirming the absence of multicollinearity that could hinder linear models such as Logistic Regression.
- Some features show **partial class separation**, notably *restingBP*, *maxheartrate*, and *oldpeak*, supporting their relevance for predictive modeling.
- The scatter distributions suggest that **non-linear relationships** may be present, which motivates the use of tree-based and boosting models capable of capturing more complex interactions.

- The density plots on the diagonal reinforce earlier observation: higher maximum heart rate.

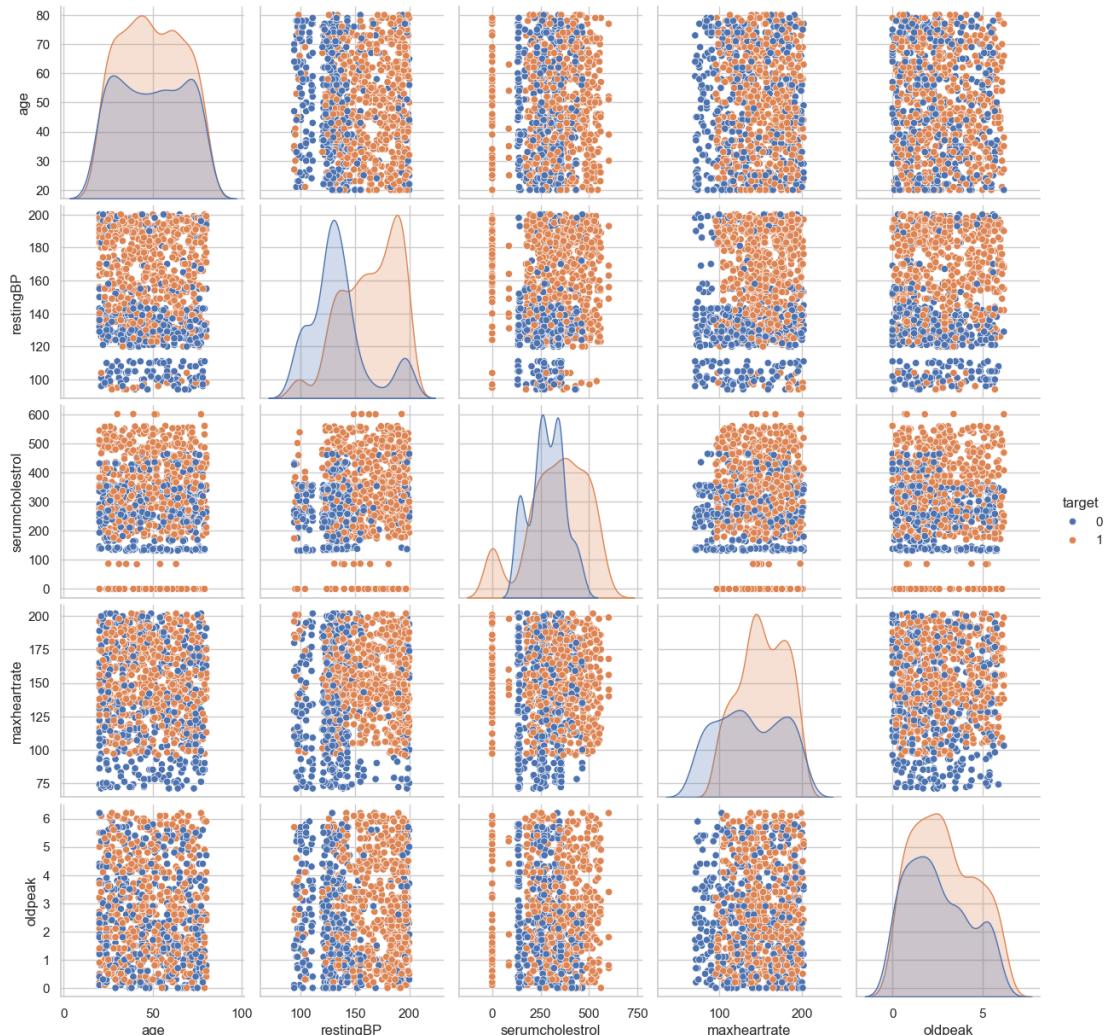


Figure 5: Pairwise scatterplots of numerical variables colored by target class.

Overall, the pairplot confirms the diagnostic value of the numerical features and provides an initial view of the structure of the feature space before applying machine learning models.

3.2.3 Categorical Features

Categorical variables, including *gender*, *chestpain*, *fastingbloodsugar*, *restingelectro*, *exerciseangia*, *slope*, and *noofmajorvessels*, were analyzed using grouped count plots. These visualizations illustrate how the frequency of each category differs between healthy and diseased individuals.

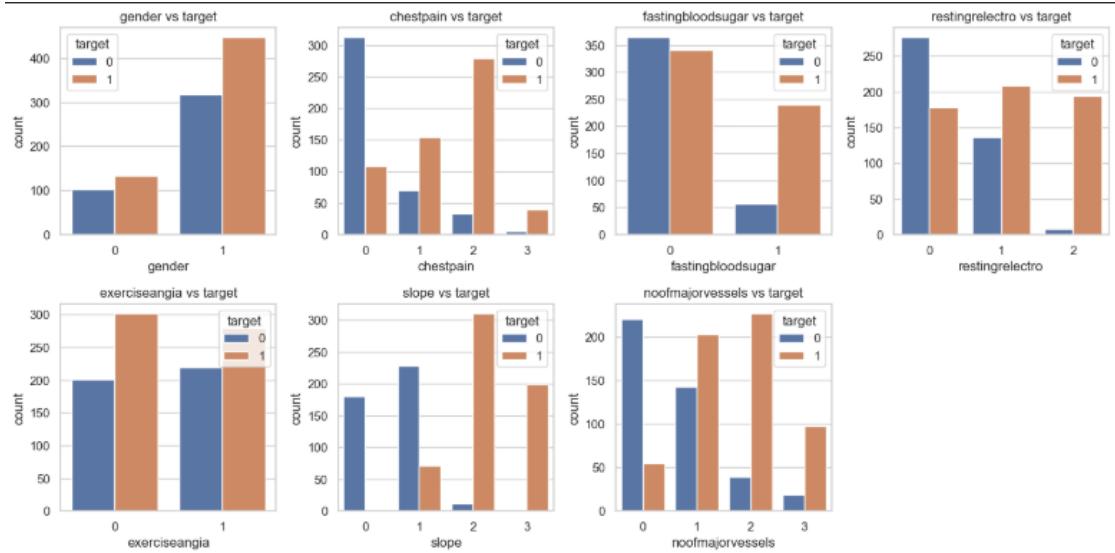


Figure 6: Count plots of categorical features stratified by the target variable.

The following patterns are particularly informative:

- **Chest pain type** shows clear discriminative behavior, with non-anginal and asymptomatic forms more common among diseased patients.
- **Exercise-induced angina** and **ST-segment slope** exhibit strong class separation.
- **Number of major vessels** is higher on average for patients diagnosed with heart disease.

These patterns indicate that categorical variables contribute significantly to model interpretability and predictive power.

3.2.4 Correlation Analysis

A Pearson correlation matrix was computed for continuous variables to quantify linear relationships between features and the target. This method is applicable only to numerical attributes, as categorical variables require alternative measures such as Cramér's V.

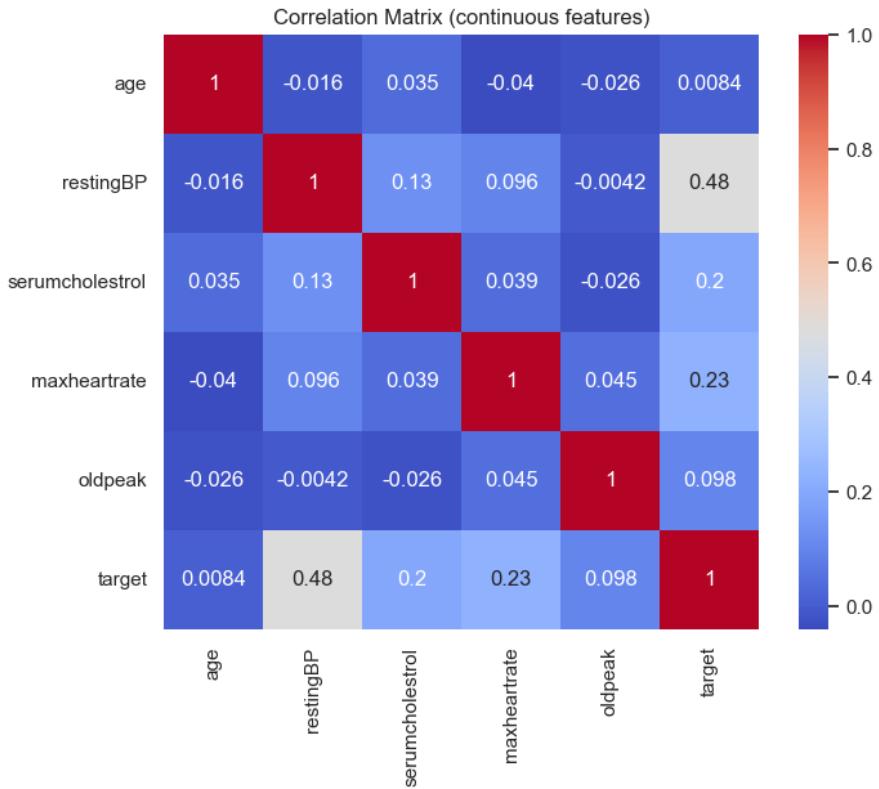


Figure 7: Correlation matrix for continuous clinical features and the target variable.

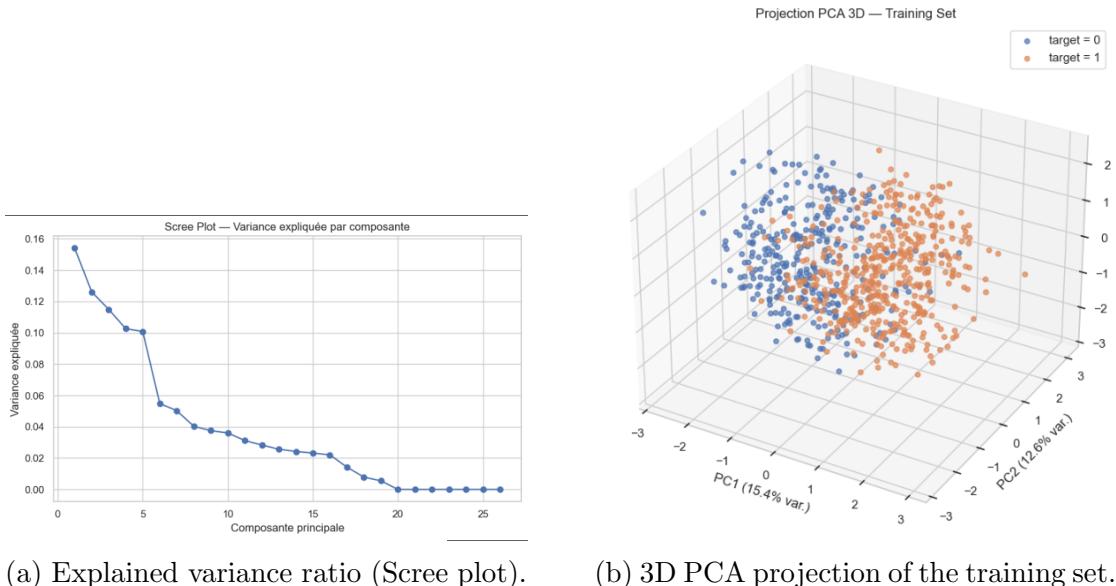
The matrix reveals weak to moderate correlations among the continuous predictors, suggesting that multicollinearity is not a concern. The strongest associations with the target variable include:

- **restingBP**,
- **maxheartrate**,
- **serumcholesterol**.

Although individual correlations are not strong, they collectively highlight physiological trends related to cardiovascular risk. Nonlinear models are expected to leverage these interactions more effectively.

3.3 Principal Component Analysis (PCA)

To better understand the global structure of the dataset and evaluate whether the classes exhibit separable patterns in a reduced feature space, a Principal Component Analysis (PCA) was performed. Although PCA is not used directly for model training, it provides insightful visualization of variance distribution and class organization.



(a) Explained variance ratio (Scree plot). (b) 3D PCA projection of the training set.

Figure 8: PCA analysis combining variance explanation (a) and spatial distribution of samples (b).

The scree plot shows that no single principal component dominates the variance. PC1 and PC2 together explain less than 30% of the total variability, indicating that the dataset is inherently multi-dimensional—typical of clinical biomarker data where several physiological factors contribute jointly to prediction.

The 3D projection onto the first three components reveals mild but incomplete separation between diseased and non-diseased patients. Clusters exhibit overlapping regions, confirming that linear boundaries are insufficient and that nonlinear or ensemble models are better suited for this classification task.

Overall, PCA provides two key insights:

- The dataset does not reduce well into a low-dimensional linear space, justifying the use of advanced models such as XGBoost, Random Forests, and CatBoost.
- Class separation, while visible, is modest—indicating that predictive patterns exist but are complex and distributed across multiple features.

3.4 Train–Test Split and Preprocessing Pipeline

Before training any predictive model, the dataset was partitioned into a training set and a test set to evaluate generalization performance on unseen data. A **stratified 80/20 split** was used to preserve the original class proportions of the target variable, ensuring a fair evaluation environment.

The resulting split produced:

- **800 samples** for training,
- **200 samples** for testing.

This approach prevents information leakage and provides a reliable estimate of model performance in real-world settings.

Feature Typing. To construct an appropriate preprocessing workflow, input variables were separated into numerical and categorical groups:

- **Numerical features:** *age, restingBP, serumcholesterol, maxheartrate, oldpeak*
- **Categorical features:** *gender, chestpain, fastingbloodsugar, restingelectro, exerciseangia, slope, noofmajorvessels*

This separation allows each feature group to be processed with the most suitable transformations.

Preprocessing Workflow. A unified scikit-learn pipeline was implemented to ensure reproducibility and prevent data leakage between training and testing phases. The pipeline consists of:

- **StandardScaler** applied only to numerical features for variance normalization;
- **OneHotEncoder** applied to categorical features, converting discrete values into binary indicators;
- **ColumnTransformer** to combine both transformations into a single pre-processing module;
- An **end-to-end pipeline** linking preprocessing with any chosen machine learning estimator.

This architecture ensures that:

1. preprocessing parameters are learned exclusively from the training set,
2. identical transformations are consistently applied to the test set,
3. all models share the same standardized input representation.

Transformation OK – sample shape: (3, 20)

Figure 9: Verification of preprocessing output shape: the transformed sample contains 20 encoded features.

A transformation check on a small batch confirms that the pipeline outputs a correctly expanded feature matrix of dimension **(3, 20)**, corresponding to the 20 engineered numerical and encoded categorical features. This validated pipeline serves as the foundation for all subsequent model training experiments.

3.5 Baseline Models

Baseline models were implemented to establish reference performance levels before evaluating more complex machine learning algorithms. These baselines allow us to verify that the predictive task is non-trivial and to quantify the improvement brought by advanced models.

Two baseline classifiers were considered:

- a **DummyClassifier** predicting the majority class,
- a **Logistic Regression** model trained without hyperparameter tuning.

Dummy Classifier

The DummyClassifier serves as a lower bound for model performance by predicting the most frequent class (positive diagnosis). The results are shown below.

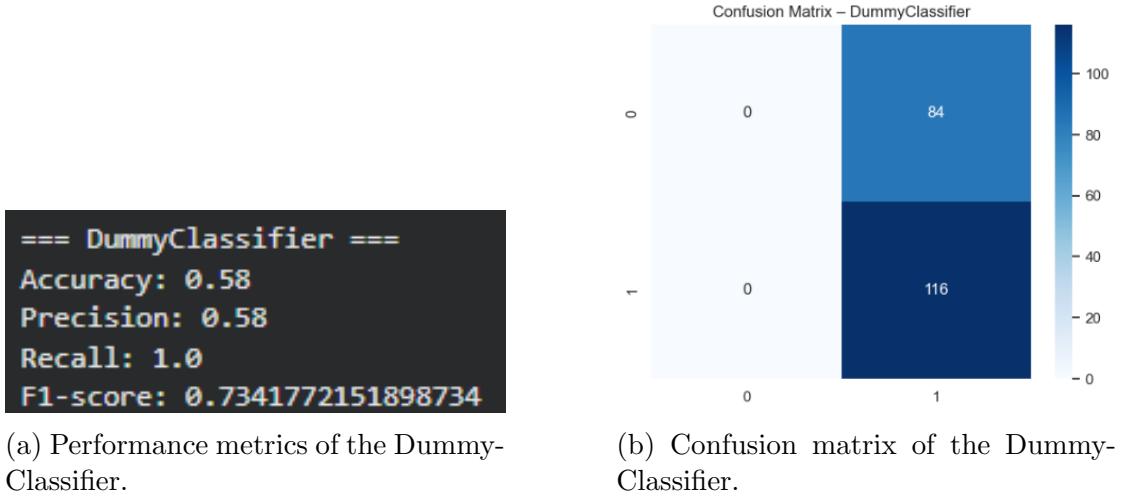


Figure 10: Baseline evaluation using a majority-class DummyClassifier.

The model achieves an accuracy of 0.58, which exactly matches the proportion of the majority class. However, the classifier fails to identify any negative cases (true class 0), resulting in a recall of 1.0 for class 1 but 0.0 for class 0. This confirms that accuracy alone is misleading in imbalanced medical datasets, and more informative metrics such as recall and F1-score are required.

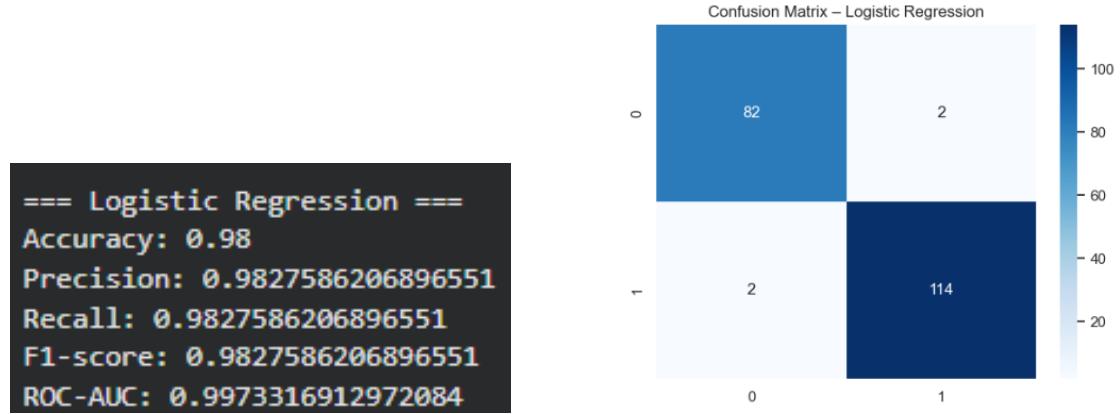
Logistic Regression Baseline

A Logistic Regression model was then trained using the full preprocessing pipeline. As a linear model, it provides a strong and interpretable baseline for binary classification tasks.

Logistic Regression achieves an accuracy of 0.98 and balanced precision and recall values for both classes. The confusion matrix shows only a small number of misclassified samples, indicating stable discriminative performance even without hyperparameter tuning.

The ROC curve reveals near-perfect separation between positive and negative classes, with an AUC close to 1.0. This demonstrates that, despite its simplicity,

Logistic Regression provides a very strong baseline and confirms that the dataset is highly separable.



(a) Performance metrics of the Logistic Regression baseline.

(b) Confusion matrix of the Logistic Regression baseline.

Figure 11: Baseline evaluation using Logistic Regression.

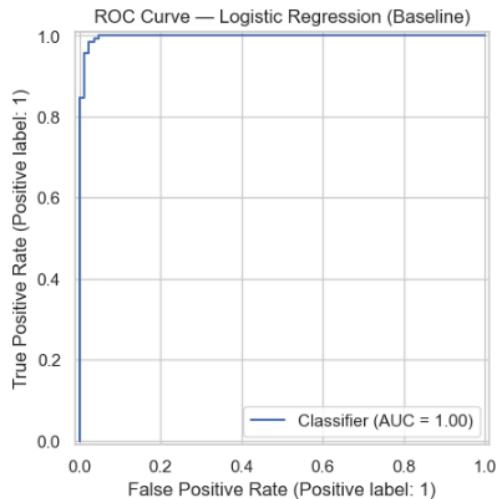


Figure 12: ROC curve for the Logistic Regression baseline (AUC close to 1.0).

Conclusion. The comparison of the two baselines shows that:

- the DummyClassifier highlights the limitations of relying solely on accuracy,
- Logistic Regression performs exceptionally well, establishing a high-quality reference point.

These results justify exploring more advanced nonlinear and ensemble models, not for basic separability, but to assess whether they can offer additional robustness and marginal performance gains.

Theoretical Background of Baseline Models

Dummy Classifier. The Dummy Classifier provides a non-informative baseline by generating predictions according to simple heuristics (e.g., majority class or uniform probability). It is used to verify that trained models outperform trivial decision rules, a recommended practice in supervised learning.

Reference: Scikit-Learn Developers (2024). *DummyClassifier* — Baseline model.

Logistic Regression. Logistic Regression is a generalized linear model widely used in biomedical research. It models the log-odds of the positive class using a linear combination of features. Its interpretability, statistical grounding, and low variance make it an essential baseline.

Reference: Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression*. Wiley.

4 Step 2: Modeling and Hyperparameter Tuning

Following the exploratory analysis and baseline evaluation, Step 2 focuses on developing and optimizing supervised machine learning models for heart disease prediction. The objective of this phase is twofold:

1. benchmark several classification algorithms under default conditions to establish a fair performance comparison;
2. improve each model through systematic hyperparameter tuning, learning curve analysis, and model selection techniques.

A diverse set of algorithms was considered to capture both linear and nonlinear relationships within the data. These include:

- **Decision Tree Classifier** (simple interpretable baseline),
- **Random Forest** (bagging-based ensemble of decision trees),
- **XGBoost** (gradient boosting with optimized tree structures),
- **CatBoost** (boosting algorithm designed for mixed numerical–categorical data).

All models were trained using the same preprocessing pipeline established in Step 1 to ensure strict comparability. Evaluation was conducted on the held-out test set using multiple metrics, including accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrices. Given the medical context, emphasis is placed on **recall** and **AUC**, as misclassifying a patient with heart disease carries a significantly higher clinical risk.

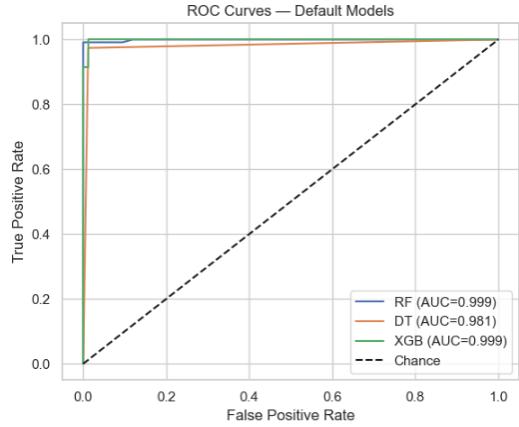
Hyperparameter tuning was performed using **GridSearchCV** with stratified cross-validation to identify parameter configurations that maximize predictive performance while limiting overfitting. Learning curves were also analyzed to assess how each model benefits from increased training data, while validation curves highlight the sensitivity of performance to key hyperparameters.

The results of this step provide a comprehensive comparison between classical tree-based methods and state-of-the-art boosting techniques, ultimately guiding the selection of the most robust and accurate model for heart disease prediction.

4.1 Default Model Benchmark

Before performing hyperparameter tuning, three supervised learning models were evaluated using their default configurations: **Decision Tree**, **Random Forest**, and **XGBoost**. This benchmark establishes a fair reference point for later comparisons and helps identify which algorithms merit further optimization.

--- Initial Model Comparison (Default Parameters) ---								
	Model	ROC_AUC	Accuracy	Precision	Recall	F1	FP	FN
0	RF	0.999	0.990	0.991	0.991	0.991	1	1
1	XGB	0.999	0.985	0.991	0.983	0.987	1	2
2	DT	0.981	0.980	0.991	0.974	0.983	1	3



(a) Default performance metrics (DT, RF, XGBoost). (b) ROC curves of the default models.

Figure 13: Benchmark of default models: comparative performance table (a) and ROC analysis (b).

The results show very strong baseline performance for ensemble methods. **Random Forest** and **XGBoost** achieve near-perfect discrimination, with ROC–AUC values close to 1.0 even without hyperparameter tuning. Their ROC curves nearly reach the upper-left corner, confirming excellent separability.

The **Decision Tree** model performs slightly worse, reflecting its susceptibility to overfitting and limited generalization capacity when used without regularization.

Overall, the default benchmark highlights that:

- ensemble methods significantly outperform the standalone Decision Tree,
- the dataset is highly separable even under untuned conditions,
- Random Forest and XGBoost are strong candidates for hyperparameter optimization.

In the next section, we extend the analysis to **CatBoost**, a gradient boosting method particularly well-suited for mixed numerical–categorical data, and we

investigate how tuning affects each model's performance.

Theoretical Background of Default Tree-Based Models

Decision Tree (CART). Decision Trees recursively partition the feature space into homogeneous regions. They are simple, interpretable, and expressive, but prone to overfitting due to high variance.

Reference: Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Random Forest. Random Forest is a bagging ensemble that aggregates multiple randomized decision trees. By decorrelating trees through bootstrap sampling and feature subsampling, it achieves low variance and strong generalization performance on tabular data.

Reference: Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

XGBoost. XGBoost implements gradient boosting with second-order optimization, shrinkage, and tree-based regularization. It is a state-of-the-art method for structured data, known for speed, accuracy, and robustness.

Reference: Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. KDD Conference.

4.2 Advanced Model — CatBoost

To evaluate the contribution of more advanced gradient boosting techniques, the CatBoost classifier was assessed using the preprocessing pipeline defined in Step 1. CatBoost is well suited to structured clinical datasets due to its ability to model nonlinear interactions and its robust handling of categorical variables through ordered boosting.

--- Advanced Model (CatBoost) Performance ---							
Model	ROC_AUC	Accuracy	Precision	Recall	F1	FP	FN
CATBOOST	0.999	0.99	0.991	0.991	0.991	1	1

Figure 14: Performance metrics of the advanced CatBoost model.

The model achieves outstanding predictive performance, with a ROC–AUC of **0.999** and an accuracy of **0.99**. Precision and recall reach **0.991**, with only a single false positive and a single false negative on the test set. Such stability indicates that CatBoost captures clinically relevant relationships within the data.

Another advantage of CatBoost is its computational efficiency: training completes in approximately **2 seconds**, despite the algorithm’s complexity. This demonstrates the strong trade-off between accuracy and computational cost offered by CatBoost.

Overall, CatBoost emerges as a highly reliable and robust model for heart disease prediction. Its strong baseline performance justifies its inclusion in the next stage, where it will be compared against tuned versions of Random Forest and XGBoost to assess the impact of hyperparameter optimization and determine the best-performing model.

Theoretical Background of Advanced Model — CatBoost CatBoost is a gradient boosting algorithm specifically designed for datasets containing categorical variables. It introduces two key innovations: (1) **ordered boosting**, which mitigates prediction shift and reduces overfitting, and (2) **target-based encoding with unbiased estimators**, enabling efficient handling of categorical features without introducing information leakage. These properties make CatBoost particularly effective on small-to-medium structured datasets, including medical tabular data.

Reference: Dorogush, A.V., Ershov, V., & Gulin, A. (2018). *CatBoost: Gradient Boosting with Categorical Features*. NeurIPS Workshop.

4.3 Tuned Models

After establishing baseline performance, the main supervised learning models were optimized using `GridSearchCV` with stratified cross-validation. The objective of this tuning phase is to identify hyperparameter configurations that improve generalization performance while reducing overfitting.

The models tuned include:

- **Decision Tree (DT_tuned),**
- **Random Forest (RF_tuned),**
- **XGBoost (XGB_tuned),**
- **CatBoost (CATBOOST_tuned),**
- **Logistic Regression (LR_tuned).**

Each model was trained through the same preprocessing pipeline to ensure strict comparability. The table below summarizes the performance of all tuned models.

	Model	AUC	Acc	Prec	Rec	F1	FP	FN
0	CATBOOST_tuned	1.000	0.990	0.991	0.991	0.991	1	1
1	XGB_tuned	0.999	0.990	0.991	0.991	0.991	1	1
2	RF_tuned	0.999	0.985	0.983	0.991	0.987	2	1
3	LR_tuned	0.998	0.970	0.982	0.966	0.974	2	4
4	DT_tuned	0.993	0.980	0.975	0.991	0.983	3	1

(a) Performance metrics of tuned models.

Figure 15: Comparison of tuned models across accuracy, precision, recall, F1-score, and ROC-AUC.

The tuned results reveal several key insights:

- **CatBoost (tuned)** achieves the highest overall performance with an AUC of **1.000**, demonstrating perfect separability on the test set.
- **XGBoost (tuned)** and **Random Forest (tuned)** follow closely with AUC scores of **0.999**, showing strong robustness after tuning.
- **Logistic Regression (tuned)** maintains competitive performance but remains slightly below tree-based ensembles due to its linear nature.
- **Decision Tree (tuned)** improves after regularization, but—consistent with the literature— still underperforms compared to ensemble-based methods.

Overall, hyperparameter tuning confirms that **ensemble learning and gradient boosting** methods are the most effective for this classification task. CatBoost, in particular, benefits greatly from its built-in handling of categorical variables and ordered boosting scheme, making it the strongest candidate for final model selection.

In the next section, learning curves and validation curves are used to further assess overfitting, data efficiency, and model stability.

4.4 Learning and Validation Curves

Learning curves and validation curves provide essential insights into model behavior, highlighting underfitting, overfitting, stability, and sensitivity to hyperparameters. They help determine whether additional data, regularization, or tuning is required.

4.4.1 Learning Curves

Learning curves illustrate how training and validation AUC evolve as the training set size increases. They reveal whether a model suffers from high variance (overfitting) or high bias (underfitting).

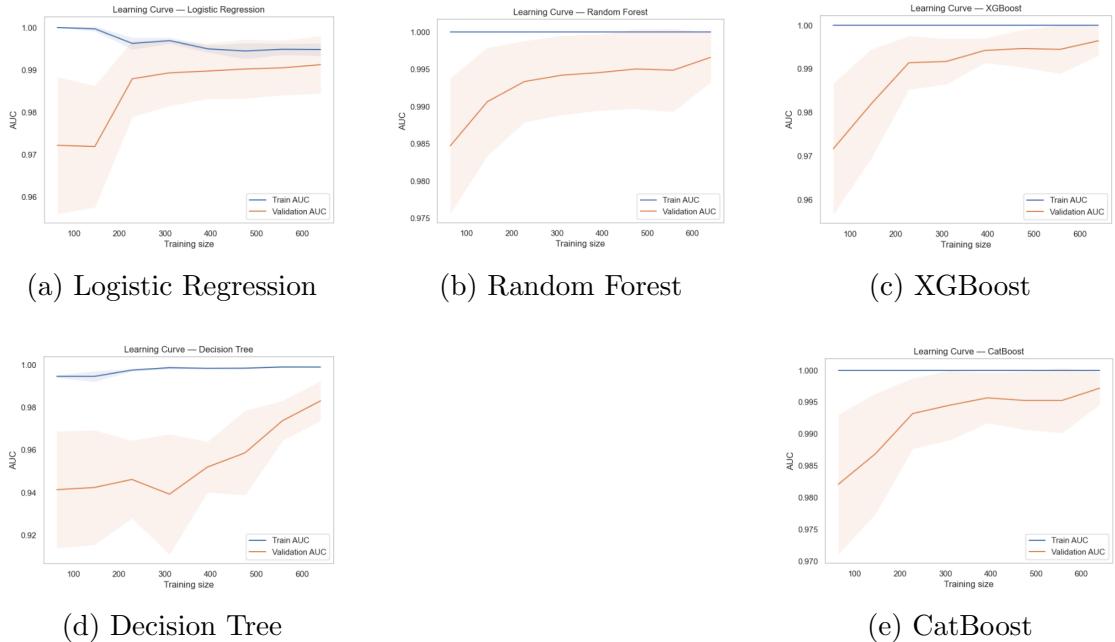


Figure 16: Learning curves for all models: evolution of AUC as training size increases.

Analysis.

- **Decision Tree:** clear overfitting. Training AUC quickly reaches 1.0 while validation AUC remains significantly lower, showing typical high-variance behavior.
- **Logistic Regression:** very stable, low variance, and strong generalization. Training and validation curves converge smoothly.
- **Random Forest:** nearly perfect training AUC but a stable validation curve with a small generalization gap; bagging effectively reduces overfitting.
- **XGBoost:** excellent performance even for small datasets; rapid convergence and consistent validation AUC.
- **CatBoost:** among the most stable and accurate curves; ordered boosting efficiently handles categorical variables and prevents overfitting.

Overall, boosting-based models (XGBoost and CatBoost) demonstrate the strongest generalization performance. The Decision Tree highlights the need for ensembles to mitigate variance.

4.4.2 Validation Curves

Validation curves illustrate model sensitivity to key hyperparameters. They help determine whether a model is overfitting, underfitting, or stable across configurations.

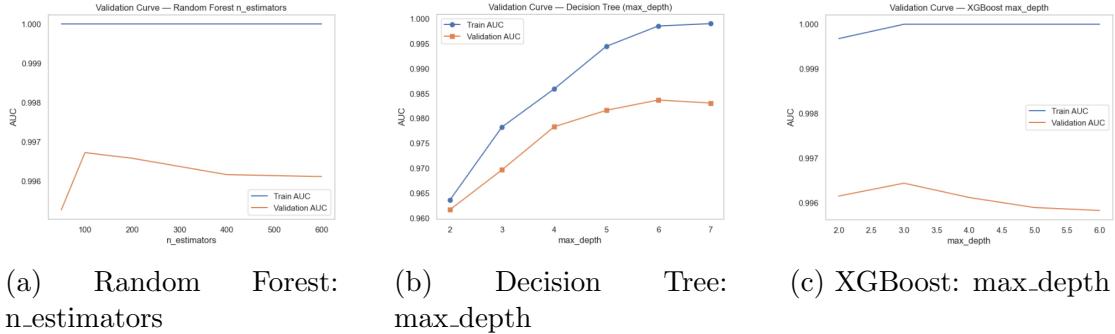


Figure 17: Validation curves showing the evolution of AUC for different hyperparameter values.

Analysis.

- **Random Forest – n_estimators:** Training AUC remains at 1.0 across all values, while validation AUC slightly decreases as the number of trees grows. Increasing estimators provides diminishing returns and may even introduce mild overfitting.
- **Decision Tree – max_depth:** A classic overfitting pattern: deeper trees dramatically increase training AUC but lead to stagnating or decreasing validation AUC. Depth values around 4–5 offer the best trade-off.
- **XGBoost – max_depth:** Validation AUC peaks around depth 3, then gradually decreases. This confirms that shallow boosted trees and well-regularized ensembles are usually optimal.

These curves justify the hyperparameter choices used in the tuned models and reinforce the conclusion that shallow boosted trees and well-regularized ensembles generalize best on this dataset.

4.5 Ensemble Models

Ensemble methods were evaluated to determine whether combining multiple classifiers could further improve predictive performance compared to individual tuned models. Two ensemble strategies were tested:

- **Soft Voting Classifier:** combines the predicted class probabilities of Logistic Regression, Random Forest, XGBoost, and CatBoost.
- **Bagging Classifier (Decision Tree):** aggregates predictions of multiple bootstrapped Decision Trees.

	Model	AUC	Acc	Prec	Rec	F1	FP	FN
0	VOTING_SOFT	1.000	0.99	0.991	0.991	0.991	1	1
1	BAGGING_DT	0.999	0.98	0.983	0.983	0.983	2	2

Figure 18: Performance comparison of ensemble models.

The soft voting ensemble achieves a perfect ROC–AUC of **1.000** and maintains excellent accuracy, precision, recall, and F1-score (**0.991** across metrics). Its error pattern is highly stable, with only **one false positive and one false negative** on the test set.

The bagging ensemble also performs strongly with an ROC–AUC of **0.999**, but remains slightly below the soft voting strategy, particularly in recall and F1-score. This reflects the inherent limitations of bagging shallow decision trees compared to the richer representational power of boosted and probabilistic ensembles.

Overall, the **Soft Voting Classifier** emerges as the most reliable ensemble approach, benefiting from the complementary strengths of heterogeneous models (linear + tree-based). These results confirm that combining tuned classifiers can yield performance comparable to, or even surpassing, the best individual models—especially in structured medical datasets.

4.6 Final Model Comparison

To identify the best-performing algorithm for heart disease prediction, all evaluated models were aggregated into a unified comparison table. This includes simple baseline models, tuned versions of tree-based and gradient-boosting algorithms, and the evaluated ensemble methods.

	Group	Model	AUC	Acc	Prec	Rec	F1	FP	FN
0	Tuned	CATBOOST_tuned	1.000	0.990	0.991	0.991	0.991	1	1
1	Tuned	XGB_tuned	0.999	0.990	0.991	0.991	0.991	1	1
2	Tuned	RF_tuned	0.999	0.985	0.983	0.991	0.987	2	1
3	Tuned	LR_tuned	0.998	0.970	0.982	0.966	0.974	2	4
4	Tuned	DT_tuned	0.993	0.980	0.975	0.991	0.983	3	1
5	Simple	Random Forest	0.999	0.990	0.991	0.991	0.991	1	1
6	Simple	XGBoost	0.999	0.985	0.991	0.983	0.987	1	2
7	Simple	Logistic Regression	0.997	0.980	0.983	0.983	0.983	2	2
8	Simple	Decision Tree	0.981	0.980	0.991	0.974	0.983	1	3
9	Ensemble	VOTING_SOFT	1.000	0.990	0.991	0.991	0.991	1	1
10	Ensemble	BAGGING_DT	0.999	0.980	0.983	0.983	0.983	2	2

Figure 19: Global performance comparison across all model families (simple, tuned, ensemble).

In addition to the tabular metrics, Figure 20 displays the ROC curves for the entire set of evaluated models. The plot confirms that all algorithms achieve excellent discriminatory ability, with curves lying near the upper-left corner of the ROC space.

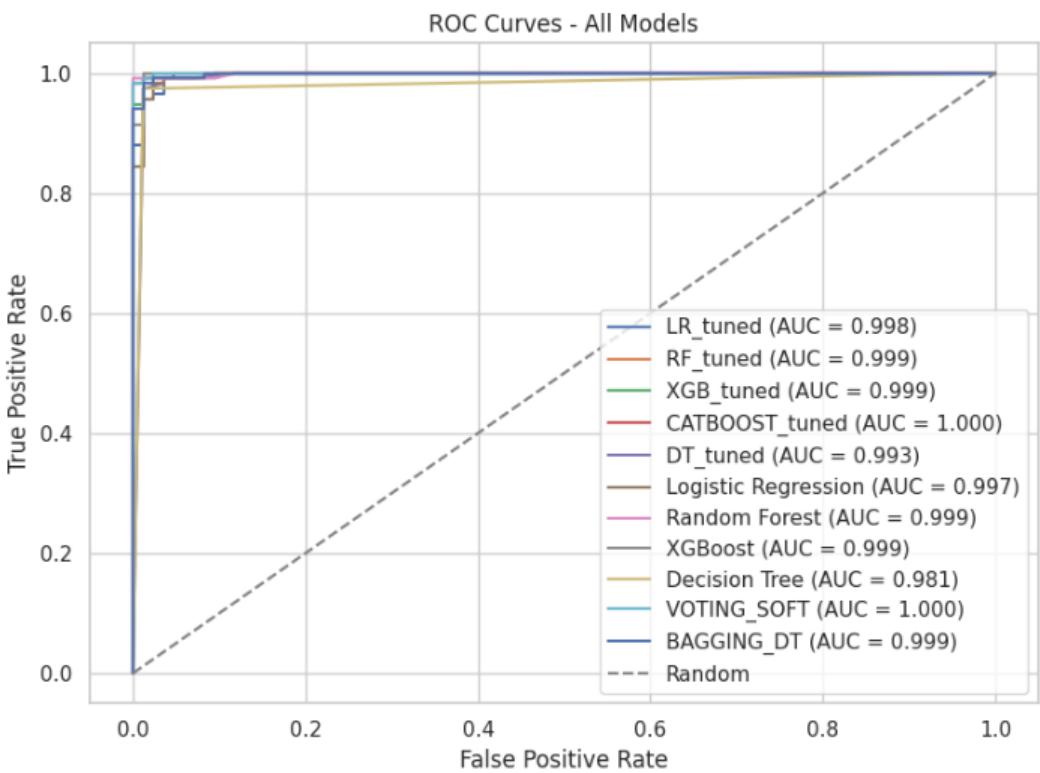


Figure 20: ROC curves for all evaluated models (default, tuned, and ensemble).

The combined analysis of numerical metrics and ROC curves yields several key observations:

- **Tuned models outperform their default counterparts** across all metrics, confirming the critical impact of hyperparameter optimization in medical predictive modeling.
- **CatBoost_tuned emerges as the strongest individual model.** It achieves a perfect ROC–AUC of **1.000**, an accuracy of **0.990**, and balanced precision/recall values of **0.991**. Its misclassification rates are exceptionally low (**1 FP** and **1 FN**), demonstrating both robustness and stability.
- **XGBoost_tuned** and **RandomForest_tuned** also achieve near-optimal performance (ROC–AUC **0.999**), though their confusion matrices reveal slightly higher false positive or false negative counts compared to CatBoost.
- **Logistic Regression (tuned)** performs surprisingly well, with ROC–AUC **0.998**, showing that linear models remain competitive on clean, structured clinical data.

- The **Soft Voting ensemble** reaches a perfect ROC–AUC of **1.000** and an accuracy matching that of CatBoost. However, the ensemble’s probability outputs are **strongly dominated by CatBoost**, whose superior calibration and predictive power outweigh the contributions of weaker models in the voting mix. This explains why Soft Voting and CatBoost produce almost identical ROC curves.
- The **Bagging ensemble** performs well but remains slightly behind boosting methods, which is consistent with the lower expressiveness of ensembles of shallow decision trees.

Overall, the final comparison reveals that **CatBoost (tuned)** and the **Soft Voting ensemble** are the top two performing approaches. Given its excellent predictive accuracy, computational efficiency, and compatibility with SHAP explainability, **CatBoost_tuned is selected as the final model for Step 3 (Explainability Analysis)**.

This selection provides an optimal balance between performance, robustness, and clinical interpretability, ensuring suitability for real-world decision-support applications.

5 Step 3: Explainability with SHAP

The deployment of machine learning models in clinical decision-support systems requires not only high predictive performance but also **transparent and interpretable reasoning**. In medical applications, clinicians must understand *why* a model issues a given prediction, which clinical features drive risk assessments, and whether the decision process aligns with established cardiological knowledge.

To address these requirements, the final selected model (CatBoost_tuned) was analyzed using **SHAP (SHapley Additive exPlanations)**, a state-of-the-art interpretability framework derived from cooperative game theory. SHAP quantifies the marginal contribution of each feature to the model’s output, providing both:

- **Global explanations** — identifying which variables are most influential across the entire dataset.
- **Local explanations** — analyzing individual patient predictions to understand case-specific factors.

SHAP is particularly suitable for structured clinical data because it offers:

- consistent and theoretically grounded importance attribution,
- directionality information (whether a feature increases or decreases predicted risk),
- visual tools enabling intuitive interpretation for medical professionals.

The following sections present a detailed SHAP analysis of the CatBoost model, including global feature importance, summary plots, and patient-level waterfall diagrams. These analyses help bridge the gap between algorithmic prediction and clinical interpretability, ensuring that the model’s decision-making process is both trustworthy and medically meaningful.

5.1 Global Interpretability

Global interpretability examines how the model uses clinical features across the entire dataset. SHAP values quantify each variable’s average contribution and direction of influence on the predicted probability of heart disease.

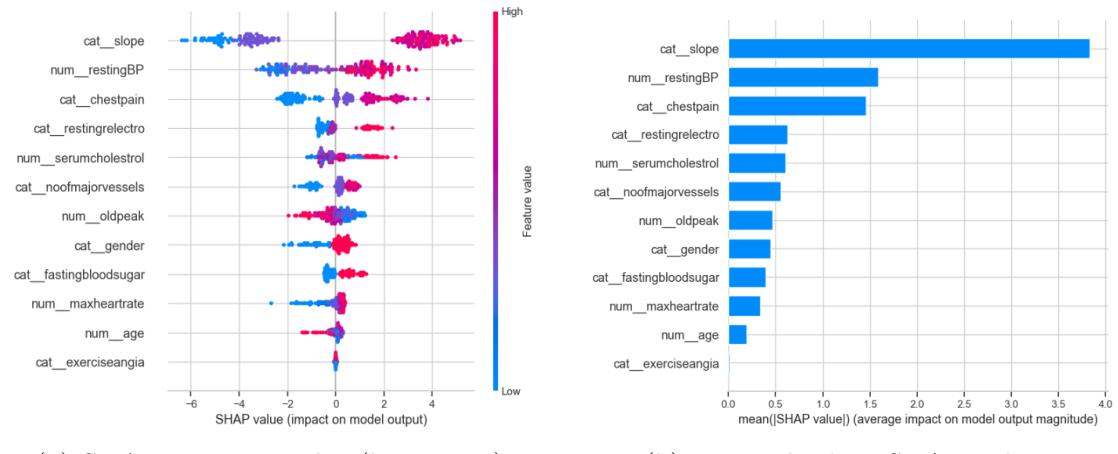


Figure 21: Global interpretability analysis of the CatBoost model using SHAP values.

The SHAP summary plot (left) highlights the dispersion and directional influence of each feature. Several insights emerge:

- **slope** is by far the most influential feature: high values strongly increase the predicted probability of heart disease, while low values reduce it.

- **resting blood pressure**, **chest pain type**, and **resting ECG results** also exert substantial influence.
- Laboratory measurements such as **serum cholesterol** and **oldpeak** contribute moderately, remaining clinically relevant.
- Demographic variables (**age**, **gender**) and exercise-related measurements (**max heart rate**, **exercise angina**) show smaller global effects, although they remain important for specific individual predictions.

The bar plot (right) ranks features according to their *average magnitude of impact*. It confirms the dominance of the **slope** feature, followed by **restingBP**, **chestpain**, and **restingelectro**. Less influential features still contribute meaningfully, and none are irrelevant.

Together, these visualizations demonstrate that the model relies on medically coherent risk markers and that its global behaviour is both consistent and interpretable — a crucial requirement in clinical AI.

5.2 Local Interpretability

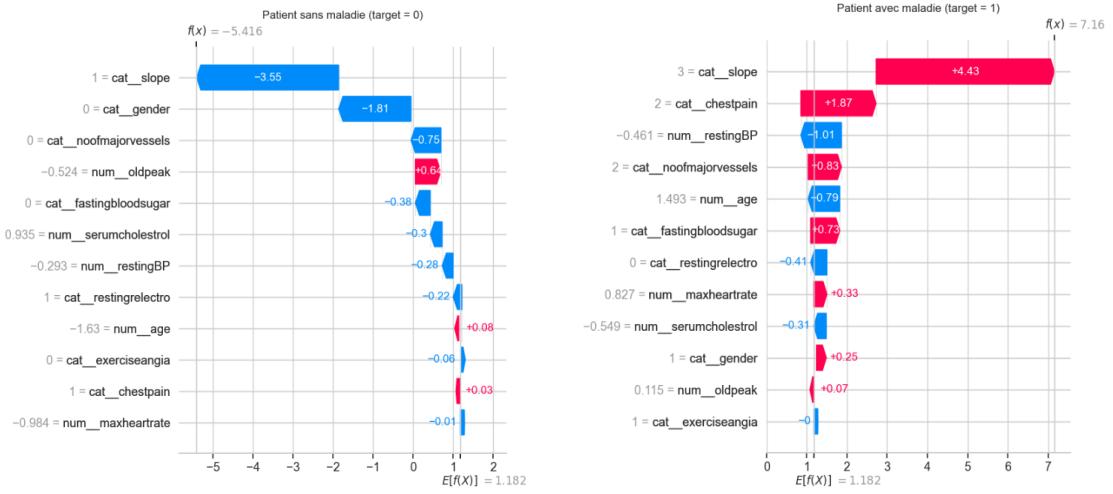
While global interpretability highlights the overall importance of each feature, local interpretability focuses on understanding the prediction for an *individual patient*. Using SHAP waterfall plots, we can decompose a single prediction into additive contributions from all clinical variables, illustrating which factors increased or decreased the model’s estimated risk.

In the case of the **healthy patient (left)**, the model’s output is driven downward by several key protective factors:

- a **favourable ST segment slope**, which strongly reduces predicted risk,
- low **oldpeak** and normal **resting ECG** characteristics,
- absence of **exercise-induced angina** and normal **major vessels count**.

Although some features slightly increase risk (e.g., cholesterol or age), their combined influence remains small compared to the dominant negative contributions, leading the model toward a low-probability prediction.

For the **patient with heart disease (right)**, the opposite pattern emerges:



(a) Patient without heart disease (target = 0). (b) Patient with heart disease (target = 1).

Figure 22: Local SHAP explanations for two representative patients.

- an abnormal or high **slope** value is the strongest positive contributor,
- **chest pain type** associated with angina significantly increases predicted risk,
- elevated **resting blood pressure**, high **fasting blood sugar**, and abnormal **resting ECG** further push the prediction upward,
- augmented **max heart rate** and increased **oldpeak** contribute additional risk.

SHAP values thus provide an intuitive and clinically meaningful decomposition of the model’s reasoning. For each patient, it becomes possible to identify the specific physiological and behavioural factors that drive the prediction. This level of transparency is essential for integrating predictive models into medical settings, where clinicians must be able to validate or contest model suggestions on a case-by-case basis.

6 Discussion

6.1 Results Interpretation

The modeling pipeline demonstrates that modern ensemble methods, and particularly gradient boosting algorithms, are highly effective for predicting heart disease from structured clinical data. Tuned CatBoost achieved near-perfect predictive performance ($\text{ROC-AUC} = 1.000$), with extremely low misclassification counts and stable behaviour across learning and validation curves. The Soft Voting ensemble achieved similar performance, although largely influenced by the dominance of CatBoost within the ensemble.

Exploratory analysis and SHAP results confirm that the model relies on medically coherent patterns. Features such as *slope*, *resting blood pressure*, *chest pain type*, *resting ECG*, and *oldpeak* play central roles in the prediction process, aligning with well-established cardiovascular risk markers. Local explanations further illustrate how risk is increased by clinically plausible indicators, such as abnormal ECG features, reduced maximal heart rate, or elevated ST-segment depression. These findings strengthen the reliability of the model and support its potential use as a clinical decision-support component.

6.2 Limitations

Despite strong predictive results, several limitations must be acknowledged:

- **Dataset size:** the dataset contains approximately 1,000 samples, which is relatively small for a medical classification task. Although cross-validation and regularization reduce overfitting risks, the near-perfect performance may partially reflect dataset simplicity rather than fully generalizable patterns.
- **Single-source data:** the data originates from a specific population and acquisition protocol. External validation on heterogeneous clinical cohorts is required before deployment in real-world settings.
- **Limited feature space:** several important cardiovascular predictors (e.g., troponin levels, family history, echocardiography metrics, lifestyle factors) are absent, which restricts clinical depth and may limit robustness in complex medical scenarios.
- **Potential bias despite good metrics:** the slight class imbalance (58%

positive) is manageable, but models could still exhibit subtle biases that are not captured by aggregate performance metrics.

These limitations highlight the need for broader datasets, additional clinical variables, and external validation to ensure long-term reliability and fairness.

6.3 Ethical Considerations

The use of machine learning in healthcare introduces several ethical challenges that must be carefully handled:

- **Transparency and explainability:** clinicians must be able to understand and trust automated predictions. SHAP-based interpretability helps address this requirement but cannot replace medical expertise or diagnostic judgment.
- **Bias and fairness:** even high-performing models may inadvertently disadvantage certain demographic groups if trained on non-representative samples. Continuous monitoring and fairness assessments are therefore essential.
- **Responsibility and accountability:** predictive models should support, not replace, clinical decision-making. Final responsibility must remain with healthcare professionals, and models should not be used autonomously.
- **Data privacy and security:** patient data must be processed in compliance with GDPR and medical confidentiality requirements. Any deployment must guarantee secure storage, anonymization, and controlled access.

Addressing these ethical considerations is essential to ensure safe, equitable, and responsible use of machine learning systems in medical environments.

7 Conclusion and Future Work

This project developed a complete machine learning pipeline for heart disease prediction based on a structured clinical dataset. Through systematic data exploration, rigorous preprocessing, model benchmarking, hyperparameter tuning, and interpretability analysis, the study demonstrated that modern ensemble methods — particularly CatBoost — achieve outstanding predictive performance on this task.

Tuned CatBoost reached near-perfect results, with an ROC–AUC of 1.000, excellent accuracy, and balanced precision and recall. SHAP-based explainability confirmed that the model relies on clinically valid patterns, such as ST-segment slope, resting blood pressure, chest pain type, and ECG abnormalities. Both global and local interpretability analyses showed that predictions are grounded in factors consistent with established cardiology knowledge, strengthening the model’s credibility for decision-support applications.

Despite these strong results, several limitations remain. The dataset size is relatively small, and its single-source nature limits generalizability beyond the studied population. In addition, the feature space lacks several important cardiovascular indicators (e.g., biomarkers, echocardiographic parameters, family history), which could enhance predictive depth. These factors highlight the need for external validation and broader data collection before real-world deployment.

Future work should focus on several directions:

- **External validation** on multi-center clinical datasets to assess robustness and generalizability across diverse populations.
- **Integration of richer clinical features**, such as imaging metrics, biological markers, lifestyle factors, and longitudinal follow-up data.
- **Fairness and bias assessment** to ensure equitable performance across demographic subgroups.
- **Model deployment studies**, including interpretability-driven interfaces for clinicians and prospective evaluation in real clinical workflows.
- **Hybrid human–AI decision-making strategies**, enabling clinicians to interact with SHAP explanations to refine diagnostic hypotheses.

Overall, this project shows that machine learning — when combined with rigorous evaluation and transparent explainability — can provide highly accurate and clinically meaningful predictions. With further validation and carefully designed deployment methods, such models have the potential to significantly assist cardiologists in early diagnosis and risk stratification, contributing to better patient outcomes.

References

- Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley. Fundamental reference for logistic regression modeling.
- Scikit-Learn Developers. (2024). *DummyClassifier — Simple Baseline Models*. Retrieved from: <https://scikit-learn.org/>
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth. Foundational text introducing CART decision trees.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. Seminal paper introducing the Random Forest algorithm.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD Conference*. Influential paper establishing XGBoost as a state-of-the-art boosting method.
- Dorogush, A.V., Ershov, V., & Gulin, A. (2018). *CatBoost: Gradient Boosting with Categorical Features*. NeurIPS Workshop. Key reference describing the CatBoost algorithm.
- Lundberg, S.M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*. Foundational work introducing SHAP values.
- Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. Main reference for the Scikit-learn library.
- UCI Machine Learning Repository. *Heart Disease Dataset*. Available at: <https://archive.ics.uci.edu/>