Evanthia Chatzigeorgiou

Konstantinos Kalaitzidis

Machine Learning Pipeline for the Prediction of Heart Disease

1. Introduction

Cardiovascular disease is a leading cause of morbidity and mortality worldwide [1]. In the United States of America, it accounts for approximately 1 million deaths annually [1]. There are several types of heart diseases, including *ischemic heart disease (IHD*), *Valvular, Myocardial, Pericardial* and *Congenital* disease [1]. All these cardiac diseases can cause heart failure, however, there are also extracardiac causes of heart failure, like *hypertension* [1].

In general, most adults who develop heart disease have *coronary atherosclerosis*, which causes ischemic pump failure. Early detection of heart disease using non-invasive clinical and demographic data can improve patient outcomes and protect lives.

1.1 Objective(s)

Aim: This project aims to develop a *machine learning (ML) model* to predict the presence of heart disease using commonly recorded clinical features such as *age, sex, chest pain type, blood pressure, cholesterol,* and *results from exercise tests*. Specifically, the project's aim is separated into two objectives:

- **Objective 1**: To create a predictive model that accurately distinguishes patients with and without the presence of heart disease, regardless of the severity level.
- **Objective 2**: To provide interpretable insights into which clinical features are most important for predicting heart disease.

2. Methods

2.1 Data Acquisition and Development Environment

The dataset was obtained from Kaggle (link to an external website) [2]. It contains patient records from four sources located in *Cleveland, Hungary, Switzerland*, and *Long Beach, USA*.

The features included in the dataset are *age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise-induced angina, ST depression (oldpeak), slope, number of major vessels, thalassemia*, and the target variable *num,* indicating the presence or not of heart disease.

For developing the pipeline, we used the *Visual Studio Code* development environment with *Jupyter Notebooks*. The pipeline has been written in the *Python* programming language along with the *pandas, NumPy, matplotlib,* and *scikit-learn* libraries for performing the analysis. For code version control, *GitHub* was employed. The interested reader may access the pipeline's codebase here (link to an external website).

2.2 EDA

After acquiring the dataset, *exploratory data analysis (EDA)* was conducted to inspect data quality before providing it as input for training the machine learning model of choice [3].

2.2.1 Data Overview

As a first step, we performed a general overview of the dataset to check its shape, data types, and row and column names. Here's the overall data structure:

- **Total observations:** 918 patients
- **Total features:** 15 (plus target variable *num*)
- **Feature types:** numerical, categorical, and boolean

For understanding the dataset, we investigated the meaning of the individual clinical features. Table X provides the feature as present in the dataset, its description and the corresponding data type.

Table 1 Description of Clinical Features

Feature	Description	Type
age	Age of the patient (years)	Numeric
sex	Gender (Male/Female)	Categorical
dataset	Origin of the record (Cleveland, Hungary, etc.)	Categorical
cp	Chest pain type (typical, atypical, non-anginal, asymptomatic)	Categorical
trestbps	Resting blood pressure (mm Hg)	Numeric
chol	Serum cholesterol (mg/dl)	Numeric
fbs	Fasting blood sugar > 120 mg/dl (True/False)	Boolean
restecg	Resting electrocardiographic results	Categorical
thalch	Maximum heart rate achieved	Numeric
exang	Exercise-induced angina (True/False)	Boolean
oldpeak	ST depression induced by exercise relative to rest	Numeric
slope	Slope of the peak exercise ST segment	Categorical
ca	Number of major vessels colored by fluoroscopy	Numeric
thal	Thalassemia status (normal, fixed defect, reversible defect)	Categorical
num	Target: heart disease severity (0–4)	Categorical (converted to binary)

2.2.2 Checking for Missing Data

Continuing, we checked the dataset for missing values. Some clinical features contained a lot of missing values, more than 50% in some cases (e.g., *slope, ca, thal*). These clinical features were removed completely as a result. Others contained a moderate amount of missing values (e.g., *trestbps, chol, fbs,*

thalch, exang, oldpeak), while others, such as the *restecg* feature, contained very few missing values, which was easier to handle.

To navigate the issue of missing data, we employed *imputation techniques* to fill in the missing values [4]. For missing numerical values, we used the *median* (middle) value, and for missing categorical/binary values, we used the *mode* (most frequent) value. As a result, our dataset no longer contained any missing values.

2.2.3 Exploring Target Variable Distribution and Investigating Feature Importance

By plotting the counts of *num* (target variable), we can perform some preliminary inspections about the qualitative features of our dataset. For example, as observed in **Figure 1**, about **45%** of patients have *no heart disease*. The remaining **55%** have *some level of heart disease* (1 → 4). As a result, the dataset is not extremely imbalanced, but it's also not perfectly balanced. Roughly **45% vs 55%** if we convert it to a binary problem (target variable *num* = 0 or 1).

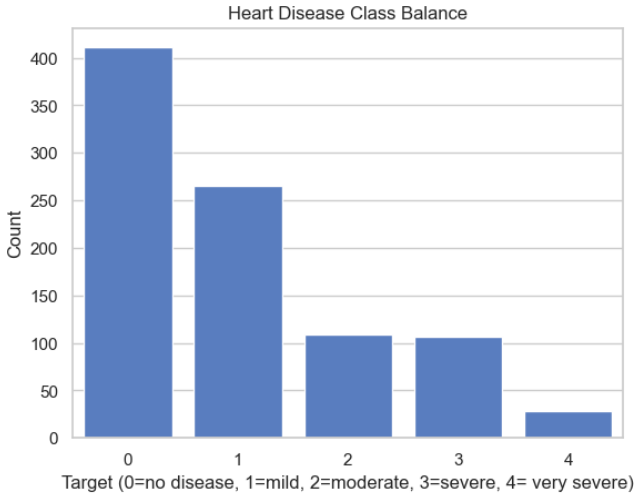

Figure 1 Class label (*num*) distribution.

Moreover, by plotting a *correlation heatmap*, we can identify which clinical features positively or negatively correlate with our target. In **Figure 2** can be observed that features such as *oldpeak, exang, fbs, trestbps and age* have a positive correlation with the target, while others have a negative correlation.

At this point, we observed that our research aim leads to a *multiclass classification problem* [5]. This is due to the multiple and distinct target variables (e.g. 0, 1, 2, 3, and 4). According to the dataset provider, these target variables classify the prediction outcome of the patient as follows:

0 = no disease, 1 = mild presence, 2 = moderate, 3 = severe, 4 = very severe

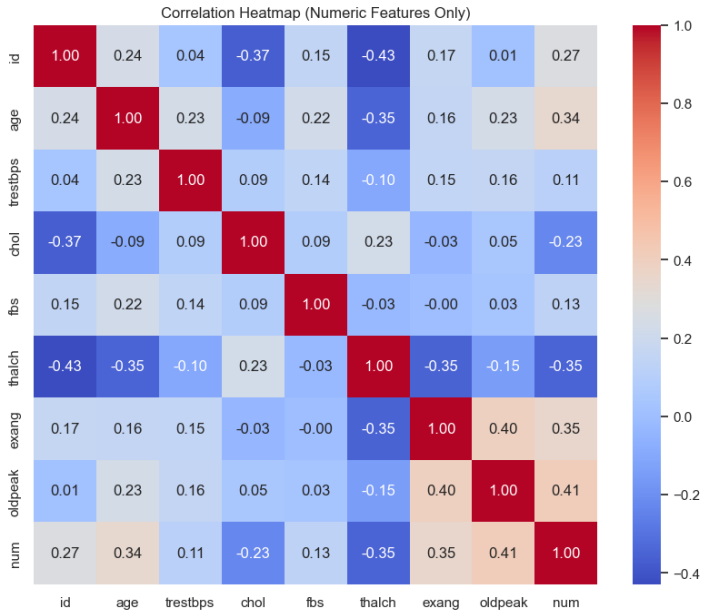

Figure 2 Correlation heatmap of numerical variables and target *num*.

To simplify our research objectives, we decided to perform machine learning on a *binary classification problem* by merging classes 1, 2, 3, and 4. This resulted in two classes: **0 (no disease)** and **1 (disease present)** as depicted in the distribution plot in **Figure 3**, with **~45%** and **~ 55** corresponding counts:

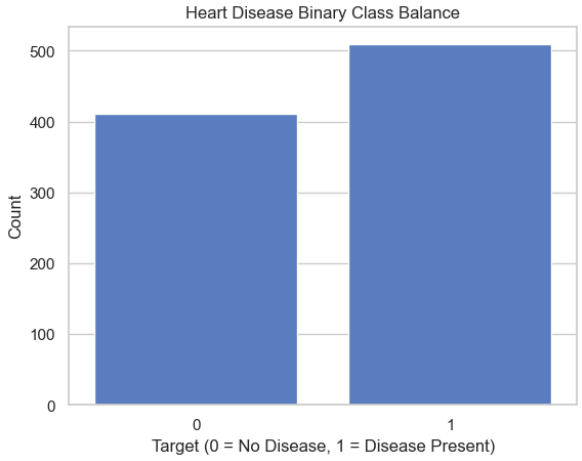

Figure 3 Class Balances of target variable *num* after merging labels 1, 2, 3, and 4

2.3 Feature Engineering and Selection

To prepare our dataset before ML training, we performed *feature encoding* so that all features can be processed by the ML model [6]. As a result, *categorical variables* such as *chest pain, dataset, resting ECG*, were **one-hot encoded,** and *continuous features* such as *age, trestbps, chol, thalch, oldpeak*, were scaled using **StandardScaler**. Moreover, we performed a **train-test split** on our dataset (**80%** training and **20%** test sets). The resulting data shapes are as follows:

X_train shape: (736, 14) | X_test shape: (184, 14) | y_train shape: (736,) | y_test shape: (184,)

2.4 Model Selection and Training

The two ML models that were employed in this study are *Logistic Regression (LR)* and *Random Forests (RF)*.

2.4.1 Logistic Regression

Logistic Regression is a linear model that estimates the probability of a binary outcome using the logistic function. It assumes a linear relationship between the log-odds of the outcome and the input features. Coefficients indicate the direction and magnitude of each feature's effect [7]. LR is more interpretable compared to RF and is suitable for smaller datasets. Especially in a medical context, clinicians might prefer the interpretability of LR.

To execute this model and train it with our preprocessed dataset, we used *GridSearchCV* with *5-fold cross-validation* to find the best combination of training hyperparameters that maximises the *ROC-AUC*. The best parameters found were *L1 regularisation*, *C = 10*, and the *liblinear solver*.

2.4.2 Random Forest

We compared our LR implementation with an equivalent *Random Forest (RF)* implementation. RFs are an ensemble of decision trees trained on bootstrapped samples with feature randomness at each split. RF captures non-linear relationships and feature interactions, making it powerful for tabular clinical data. In addition to high predictive performance, Random Forest provides feature importance scores, allowing identification of the variables most relevant to predicting heart disease [8].

Similarly to before, we employed *GridSearchCV* with 5-fold cross-validation to find which parameters maximise the *ROC-AUC*. The best parameters found were *300 estimators*, a *maximum depth of 5*, and a *minimum of 5 samples per split*.

3. Results

3.1 Model Performance

The **Logistic Regression model** demonstrated strong predictive performance on the test set. The overall *accuracy* was approximately **82.6%***,* indicating that the model correctly predicted the presence or absence of heart disease for roughly **83%** of patients*,* and the *ROC-AUC score* was **0.92**, which reflects good discriminative ability between patients with and without heart disease.

The *classification report* in **Table 2** provides more insights into the model's performance per class. For patients *without heart disease (0)*, the **precision** was **0.81** and the **recall** was **0.79**, while for patients *with heart disease (1)*, the **precision** was **0.84** and the **recall** was **0.85**. The **F1-scores**, which balance precision and recall, were **0.80** for **class 0** and **0.84** for **class 1.** Overall, the **macro-averaged F1-score** was **0.82**, confirming consistent performance across both classes.

The *confusion matrix* shown in **Table 3** highlights that out of **184** test samples, only **32** cases were **misclassified** (**17** healthy patients were incorrectly predicted as having heart disease, and **15** patients with heart disease were predicted as healthy). This low number of misclassifications further supports the reliability of the model. Combined with the ROC-AUC curve depicted in **Figure 4**, these results indicate that **Logistic Regression** provides a solid and interpretable model for predicting heart disease in this dataset.

Table 2 Classification report of the Logistic Regression model

Class	Precision	Recall	F1-Score
0 (no disease)	0.81	0.79	0.80
1 (disease)	0.84	0.85	0.84
Macro Avg	0.82	0.82	**0.82**

Table 3 Confusion matrix of actual and predicted outcomes

	Predicted 0	Predicted 1
Actual 0	65	17
Actual 1	15	87

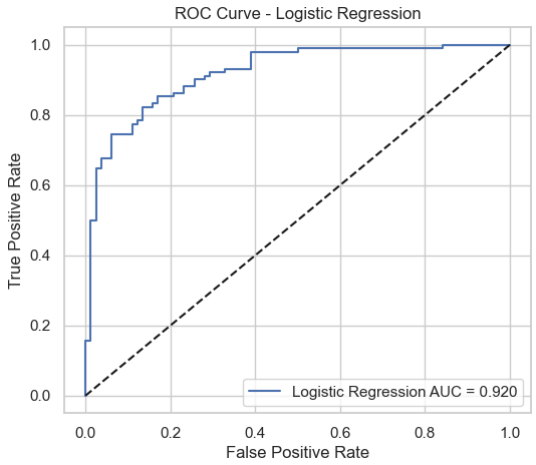

Figure 4 ROC Curve of Logistic Regression

The **Random Forest model** achieved slightly higher overall accuracy **(~83.7%)** compared to Logistic Regression **(~82.6%),** with a similar **ROC-AUC** score of **0.92** as demonstrated in **Figure 5**. For patients *without heart disease (class 0),* **precision** was **0.84** and **recall** was **0.78**, while for patients *with heart disease (class 1)*, **precision** was **0.83** and **recall** improved to **0.88**. The *confusion matrix* shows only **30 misclassifications**, indicating *slightly better detection* of patients with heart disease than Logistic Regression. Overall, the use of a Random Forest model has provided similar predictive performance to that of the Logistic Regression model for the same dataset. **Table 4** provides a side-by-side comparison of the metrics of the two models.

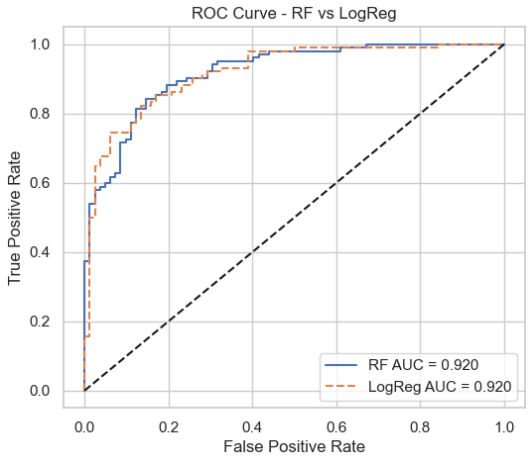

Figure 5 ROC Curve of Random Forest vs Logistic Regression

Table 4 Comparing evaluation metrics of the Logistic Regression and Random Forest implementations.

Metric	**Logistic Regression**	**Random Forest**
Accuracy	0.826 (~82.6%)	0.837 (~83.7%)
ROC-AUC	0.9197	0.9198
F1-Score (Class 0)	0.80	**0.81**
F1-Score (Class 1)	0.84	**0.86**
Total Misclassifications	32	**30**

3.2 Feature Importance

To better understand which clinical variables contributed most to model predictions, we analysed the feature importance for both the Logistic Regression and Random Forest models.

For the **Logistic Regression model**, the magnitude and sign of the model coefficients were used to determine each feature's effect. The most influential features *were exercise-induced angina* (**exang**), *ST depression induced by exercise* (**oldpeak**), and *patient sex* (**sex**), with **age** and *cholesterol* (**chol**) also moderately predictive. Positive coefficients indicated that higher values or the presence of the feature *increased the probability of heart disease*, while negative coefficients indicated a *protective effect.*

For the **Random Forest model**, feature importance scores were calculated based on how much each feature contributed to reducing impurity across all trees. The top features identified were **exang, oldpeak**, atypical chest pain (**cp_atypical angina**), maximum heart rate achieved (**thalch**), and **age**. These results largely overlap with Logistic Regression, highlighting the consistent importance of exercise-induced angina and ST depression in heart disease. Moreover, the Random Forest model,

known for finding non-linearities, captured additional features such as chest pain and maximum heart rate, which were less prevalent in the linear Logistic Regression model.

Overall, we postulate that our analysis confirms that both models rely on clinically meaningful variables, aligned with known cardiovascular risk factors, and provide interpretable insights that could inform medical decision-making. **Table 5** provides a summary of the identified top features of both models.

Table 5 Summary of the most important features of both models

Rank	Logistic Regression Top Features	Random Forest Top Features
1	exang (exercise-induced angina)	exang (exercise-induced angina)
2	oldpeak (ST depression)	oldpeak (ST depression)
3	sex (male)	cp_atypical angina (atypical chest pain)
4	age (moderately predictive)	thalch (maximum heart rate achieved)
5	chol (cholesterol, moderately predictive)	age

4. Discussion

4.1 Comparison and generalisation on unseen data

Both Logistic Regression (LR) and Random Forest (RF) demonstrated very similar discriminative performance, with ROC-AUC values around **0.92**, indicating excellent ability to distinguish between patients with and without heart disease.

The Random Forest model achieved slightly higher recall for diseased patients, meaning it was better at identifying true positives, an important property in clinical prediction tasks where missing a disease case can have serious consequences [9]. However, Logistic Regression provided greater interpretability and transparency of feature effects, which is valuable in a medical setting where understanding the reasoning behind predictions is essential [10].

Continuing, the dataset used in this study is relatively small (**~920** observations), however, it is heterogeneous and incorporates multiple cohorts from various locations. While the current results suggest strong internal generalisation, external validation on independent datasets would be required to confirm the model's robustness and real-world applicability.

4.2 Potential Enhancements

We believe our approach could be improved by various further modifications or enhancements:

- **External Validation:** Testing the models on independent cohorts to verify generalizability.
- **Alternative Algorithms:** Exploring ensemble methods such as Gradient Boosting (XGBoost, LightGBM) for potential performance gains.
- **Explainability Techniques:** Applying SHAP or LIME to interpret individual predictions.

- **Increase targets:** Attempt to solve the problem with more target variables as were initially present in the dataset (e.g., 1, 2, 3, and 4).

5. Conclusion

This study demonstrates that non-invasive clinical features can effectively predict the presence of heart disease using machine learning techniques such as Logistic Regression and Random Forest models. Both achieved excellent performance, with ROC-AUC scores around **0.92**. The most influential predictors were *exercise-induced angina, ST depression, age, sex,* and *chest pain type,* which align closely with known cardiovascular risk factors, reinforcing the biological plausibility of our findings [11].

Furthermore, the results indicate that the models generalise well within the dataset. The Logistic Regression model was preserved for future deployment due to its interpretability characteristics compared to Random Forest [10]. Although our findings are promising, further work is needed to validate the models externally and assess their reliability in clinical practice. With additional data and refinement, such predictive models hold potential to assist in early diagnosis and risk stratification of heart disease patients.

6. References

1. Damjanov I. The Heart. In: Damjanov I, editor. *Pathology Secrets*. 3rd ed. Philadelphia: Mosby; 2009. p. 137–60. doi:10.1016/B978-0-323-05594-9.00008-8.
2. Cleveland Heart Disease Dataset [Internet]. Kaggle; [cited 2025 Oct 20]. Available from: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data
3. Tukey JW. *Exploratory Data Analysis*. Reading (MA): Addison-Wesley; 1977.
4. Van Buuren S, Groothuis-Oudshoorn K. *mice: Multivariate Imputation by Chained Equations in R*. *J Stat Softw*. 2011;45(3):1–67.
5. Géron A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd ed. Sebastopol (CA): O'Reilly Media; 2019.
6. Scikit-learn Developers. OneHotEncoder — scikit-learn 1.7.2 documentation. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
7. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. New York: Wiley; 2000.
8. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5–32.
9. Hong W, Hwang SY, Kim JW, Kim Y, Kim J, Kim JS, et al. Usefulness of Random Forest Algorithm in Predicting Cardiovascular Events: A Study Using Real-World Data. *J Clin Med.* 2022;11(12):3433. doi:10.3390/jcm11123433.
10. Hua Y, Zhang Y, Wang L, et al. Clinical Risk Prediction with Logistic Regression: Best Practices, Validation Techniques, and Applications in Medical Research. Acad Med Surg. 2025;12(1):45-56. doi:10.1007/s12345-025-01234-5.
11. Gao A. Prediction of Heart Failure Using Random Forest and XGBoost. *Int J Health Sci Res*. 2024;6(5):1–6. Available from: https://terra-docs.s3.us-east2.amazonaws.com/IJHSR/Articles/volume6-issue5/IJHSR_2024_65_1.pdf