Machine Learning Project
Heart Failure Prediction

Michele Cucchiaro
Università di Bologna
Digital Transformation Management

***Abstract*—This project investigates the application of machine learning techniques to predict heart disease using a structured data-mining workflow. A dataset of 918 patient records, each containing 11 clinically relevant features, is analyzed through a pipeline including data understanding, preprocessing, feature engineering, model development, and evaluation. Three main classification algorithms—Logistic Regression, K-Nearest Neighbors, and Random Forest—are implemented and compared. Additional optimization is performed through AutoML, cross-validation, grid search, and feature selection methods. Results show that Random Forest achieves the highest accuracy (87.5%), with feature importance aligning with established medical knowledge. The study highlights the potential of machine learning as a decision-support tool in cardiovascular risk assessment.**

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, responsible for approximately 17.9 million deaths each year, representing about 31% of all global deaths. Nearly 80% of CVD-related fatalities are caused by heart attacks and strokes, with close to one-third of these deaths occurring prematurely among individuals under the age of 70. Due to their high prevalence and severe impact on public health, early detection and accurate risk prediction of cardiovascular conditions are of critical importance.

The workflow of this project was structured into several well-defined stages, namely data exploration, feature engineering, model development, and model evaluation. This modular organization allows for a systematic and reproducible approach to the analysis, ensuring that each phase contributes effectively to the overall performance of the predictive models.

The first stage focuses on data cleaning and preprocessing, where missing values, inconsistencies, and potential outliers are addressed to improve data quality. The second stage involves exploratory data analysis (EDA) and data visualization, which aim to identify underlying patterns, correlations between features, and potential predictive variables. These insights guide the subsequent model development phase, where multiple classification algorithms are trained and tuned. Finally, the evaluation stage assesses model performance using appropriate metrics, enabling a fair comparison between different approaches.

The project makes use of a publicly available dataset consisting of 918 patient records, each described by 11 clinically relevant features related to cardiovascular health. By experimenting with multiple classification techniques, the objective is to identify a model that achieves a suitable balance between predictive accuracy, interpretability, and robustness in estimating the risk of heart failure. Such a balance is particularly important in medical applications, where transparency and reliability are essential for clinical adoption.

Regarding the dataset sources, the final dataset was constructed by merging five distinct datasets that have been widely used in previous cardiovascular research studies. At present, this combined dataset represents the largest publicly available collection suitable for this type of analysis. The provenance of the individual datasets is summarized below:

- **Cleveland:** 303 observations
- **Hungary:** 294 observations
- **Switzerland:** 123 observations
- **Long Beach VA:** 200 observations
- **Stalog (Heart) dataset:** 270 observations

Although the combined datasets originally contained a total of 1,190 observations, several duplicate records were present across two or more sources. After identifying and removing 272 duplicate entries, the dataset was reduced to its final size of 918 unique observations. This deduplication process was essential to prevent bias and ensure the reliability of the machine learning models trained on the data.

2. Related work

Generally speaking, machine learning is widely used in disease diagnosis, and its presence is growing continuously, as it can analyse large and complex medical data much faster and precisely than traditional methods. The upsides of using machine learning in this field is that it can greatly help doctors in detecting diseases and generating diagnoses in a faster and more precise fashion. Considering more specifically this dataset, there have been numerous instances of machine learning projects carried out on it. The results of these projects vary, but the more popular and successful ones have achieved an accuracy of more than 90% accuracy, when consieding the precision of the prediction of the presence of an heart disease. One specific project, the more popular one

related to this dataset present of the kaggle platform, has been used as a blueprint for this work, integrating it with the workfolow presented during the lectures of this course. The specific sections will be described in detail in the following chapter.

3. Proposed Method

The methodology adopted in this project follows a structured and reproducible data–mining workflow inspired by the CRISP-DM framework. The process is divided into four main phases: data understanding, data preparation, modeling, and evaluation. Each phase is designed to progressively refine the dataset and improve the predictive performance of the machine learning models.

3.1. Data Understanding

The dataset consists of 918 patient records and 12 variables, including 11 clinical features and one binary target variable (*HeartDisease*). The target distribution is relatively balanced, with 508 positive cases (55%) and 410 negative cases (45%). This balance allows the use of standard classification algorithms without requiring resampling techniques.

The features include both numerical variables (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) and categorical variables (Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope). Initial exploratory data analysis (EDA) was performed using descriptive statistics, histograms, boxplots, and countplots. This analysis revealed several important characteristics:

- The *Cholesterol* feature contains approximately 170 zero values, which are physiologically implausible and likely represent missing data.
- The distribution of categorical variables is uneven, with some categories (e.g., ChestPainType_TA) being significantly underrepresented.
- Correlation analysis using Pearson's coefficient showed that no pair of features exceeded an absolute correlation of 0.6, indicating that multicollinearity is not a major concern.

These insights guided the subsequent preprocessing steps and informed the selection of appropriate modeling techniques.

3.2. Data Preparation

Data preparation focused on cleaning, encoding, and transforming the dataset to ensure compatibility with machine learning algorithms.

3.2.1. Handling Invalid Cholesterol Values. Since zero cholesterol values are not medically meaningful, they were replaced using mean imputation. The mean was computed only from non-zero entries to avoid bias. This approach preserves the dataset size while correcting erroneous measurements.

3.2.2. Encoding Categorical Variables. Categorical features were encoded as follows:

- **Binary encoding**: *Sex* (M=1, F=0) and *ExerciseAngina* (Y=1, N=0).
- **One-hot encoding**: applied to *ChestPainType*, *RestingECG*, and *ST_Slope*, generating a total of 10 dummy variables.

This transformation ensures that all features are numerical and suitable for algorithms such as Logistic Regression, KNN, and Random Forest.

3.2.3. Feature Selection. A manually curated list of predictive features was initially defined based on clinical relevance and EDA insights. To validate this selection, two feature selection techniques were applied:

- **RFECV (Recursive Feature Elimination with Cross-Validation)** using a Random Forest estimator.
- **RFE (Recursive Feature Elimination)** with a fixed number of selected features.

Both methods confirmed the importance of features such as Age, Cholesterol, Oldpeak, ExerciseAngina, and ST_Slope_Up, aligning with medical literature and the model's internal feature importance.

3.2.4. Normalization. For distance-based models (e.g., KNN) and PCA visualization, z-score normalization was applied. This step standardizes the scale of numerical features, preventing models from being biased toward variables with larger magnitudes.

3.3. Modeling

Three primary machine learning models were implemented and evaluated: Logistic Regression, K-Nearest Neighbors, and Random Forest. Additional optimization was performed using AutoML, grid search, and cross-validation.

3.3.1. Logistic Regression. Logistic Regression serves as a strong baseline for binary classification. It models the log-odds of the target variable as a linear combination of the input features. The model was trained using default hyperparameters and achieved an accuracy of 85.9%. Despite its simplicity, it provides interpretable coefficients and establishes a reference point for more complex models.

3.3.2. K-Nearest Neighbors. KNN is a non-parametric, instance-based classifier that predicts the class of a sample based on the majority label among its k nearest neighbors. The model was evaluated for k values ranging from 1 to 29. Accuracy peaked just below 86%, with diminishing returns for larger values of k. A confusion matrix was generated for $k = 1$ to analyze misclassification patterns. Although KNN performed reasonably well, its sensitivity to feature scaling and computational cost on larger datasets limit its applicability.

3.3.3. Random Forest. Random Forest, an ensemble of decision trees, demonstrated the best performance with an accuracy of 87.5%. The model benefits from reduced variance, robustness to noise, and the ability to capture nonlinear relationships. Feature importance analysis revealed that ST_Slope_Up, Cholesterol, Oldpeak, Age, and ChestPainType_ASY were the most influential predictors. These findings are consistent with established cardiovascular risk factors.

3.3.4. Additional Optimization. To further refine model performance, several advanced techniques were employed:

- **AutoML (FLAML)** automatically explored multiple algorithms and hyperparameters within a time budget.
- **Grid Search** was applied to KNN and Random Forest to identify optimal hyperparameters.
- **Cross-validation** (5-fold) was used to assess model generalization and reduce overfitting.
- **SVM Pipeline**: A pipeline combining StandardScaler and SVC was tuned using GridSearchCV, exploring different kernels and regularization strengths.

These methods provided additional insights into model behavior and confirmed the robustness of the Random Forest classifier.

4. Results

This section presents the performance of the machine learning models developed in the project, along with additional analyses aimed at understanding model behavior, feature relevance, and generalization capability. The evaluation focuses on accuracy, confusion matrices, feature importance, and cross-validation results. All experiments were conducted using the preprocessed dataset described in the previous section.

4.1. Model Performance Comparison

Table 1 summarizes the accuracy achieved by each of the primary models: Logistic Regression, K-Nearest Neighbors, and Random Forest. Accuracy was computed on the held-out test set obtained through a stratified train-test split.

TABLE 1. MODEL ACCURACY COMPARISON

Model	Accuracy
Logistic Regression	0.858
KNN	$\sim$0.86
Random Forest	**0.875**
AutoML (FLAML)	Comparable to RF

Random Forest achieved the highest accuracy at 87.5%, outperforming both Logistic Regression and KNN. This result is consistent with the model's ability to capture nonlinear relationships and interactions between features, which are common in medical datasets.

4.2. KNN Evaluation

The KNN classifier was evaluated for values of k ranging from 1 to 29. The accuracy curve showed an initial improvement as k increased, stabilizing around $k = 10$. Very small values of k (e.g., $k = 1$) led to overfitting, as reflected in the confusion matrix for $k = 1$, which showed a higher number of misclassifications for the negative class.

Although KNN performed reasonably well, its sensitivity to feature scaling and the curse of dimensionality limited its performance relative to Random Forest.

4.3. Random Forest Feature Importance

Random Forest not only achieved the best accuracy but also provided valuable insights into feature relevance. The top five most important features were:

- ST_Slope_Up
- Cholesterol
- Oldpeak
- Age
- ChestPainType_ASY

These results align with established medical knowledge. For example, ST-segment slope and exercise-induced angina are well-known indicators of cardiac stress, while cholesterol and age are classical cardiovascular risk factors. The consistency between model-derived importance and clinical expectations increases confidence in the model's interpretability.

4.4. Cross-Validation and Generalization

To assess generalization performance, 5-fold cross-validation was applied to the Random Forest classifier. The resulting accuracy scores showed low variance across folds, indicating that the model is stable and not overly sensitive to the specific train-test split. This reinforces the reliability of the Random Forest model as the best-performing classifier in this study.

4.5. Hyperparameter Optimization

Grid search was applied to both KNN and Random Forest to identify optimal hyperparameters. For KNN, the search explored different values of k and leaf sizes, while for Random Forest, the number of estimators was tuned. Although minor improvements were observed, the default Random Forest configuration already performed near optimally.

Additionally, AutoML (FLAML) was used to automatically explore a broader search space of algorithms and hyperparameters. The best-performing AutoML models achieved accuracy comparable to Random Forest, further validating the strength of ensemble-based approaches for this dataset.

5. Conclusions

This project explored the application of machine learning techniques to the prediction of heart disease using a structured and reproducible data-mining workflow. By integrating data understanding, preprocessing, feature engineering, model development, and evaluation, the study demonstrated how classical machine learning algorithms can effectively support clinical decision-making.

Among the models evaluated, Random Forest achieved the highest accuracy (87.5%), outperforming both Logistic Regression and K-Nearest Neighbors. Its ability to capture nonlinear relationships and provide interpretable feature importance made it particularly suitable for this task. The most influential features identified by the model—such as ST_Slope_Up, Cholesterol, Oldpeak, Age, and ChestPainType_ASY—are consistent with established cardiovascular risk factors, reinforcing the clinical validity of the results.

The use of additional techniques such as AutoML, grid search, cross-validation, and feature selection further strengthened the robustness of the findings. These methods confirmed that ensemble-based models are well-suited for heterogeneous medical datasets and that careful preprocessing, especially of categorical variables and invalid measurements, is essential for achieving reliable performance.

Overall, the study highlights the potential of machine learning as a valuable decision-support tool in healthcare. While the results are promising, future work could explore larger and more diverse datasets, incorporate additional clinical variables, or investigate advanced models such as gradient boosting or deep learning. Such extensions may further improve predictive accuracy and enhance the applicability of machine learning in real-world clinical environments.

References

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