

Classifying patients at risk for heart failure using clinical data

Mara Sánchez, Eric Yang & Omar Ramos

2025-11-21

Table of contents

Summary	1
Introduction	2
Methods	2
Results & Discussion	3
Value Ranges for Data Validation	3
Models Evaluation	6
Evaluation on Test Set	8
References	9

Summary

In this analysis, we explored various classification models with the intent of predicting whether a patient is at risk of heart failure based on clinical data and lifestyle factors of individuals. After evaluating multiple models through cross-validation, we selected Logistic Regression as our final model due to its overall superior performance across classification metrics. The model demonstrated promising results on the unseen test set, with an accuracy of 0.88 and F1-scores of 0.87 for the positive class (at risk) and 0.88 for the negative class (not at risk). From the 220 observations in the test set, the model correctly identified 97 cases at risk and 96 not at risk, reporting 16 false positives and 11 false negatives (cases predicted as not at risk when there is risk). Although the scores are encouraging for a first iteration, there is room for improvement to optimize the hyperparameters and the model's threshold settings to minimize false negative

cases, which are critical in medical applications. Overall, this model shows potential to support clinical professionals in the assessment of patients during screening.

Introduction

Cardiovascular diseases (CVDs) represent the leading cause of death worldwide, responsible for an estimate of 19.8 million deaths in 2022 and accounting for one-third of all deaths globally in people under the age of 70 (World Health Organization 2024). Most CVDs have proven to have a big correlation with an individual’s behavior, habits and environment (American Heart Association 2025). Heart Failure (HF) is a multi-faceted and life-threatening syndrome where the heart’s ability to pump and/or fill with blood is reduced, it is estimated to affect more than 64 million people globally (Savarese, Lund, and Becher 2023). Risk factors such as high blood pressure, high blood glucose levels or pre-existing diseases require early intervention to avoid the risk of developing heart failure and other complications. Since many of these risk factors are clinically measurable and changeable, early intervention is highly possible and a critical opportunity in contributing to a patient’s well-being.

This issue leads to the question of whether a machine learning (ML) model could reliably classify patients as ‘at-risk’ or ‘not-at-risk’ for heart failure based on clinical and lifestyle features. The study of this question is important because traditional risk assessment methods tend to overlook the variability and the complexity of the risk factors, meaning that subjectivity from healthcare professionals could also impact the outcome of the assessment (Barnett, Koppes, and et al. 2020). Missing the early detection of risk of HF could lead to chronic and progressive conditions associated with increase in mortality and decrease in quality of life. Additionally, ML algorithms could offer significant advantages in predicting risk of HF, when integrating it to clinical practice it could allow for a more personalized treatment for the patient (Kokori et al. 2025).

Methods

The [dataset](#) used in this project is pulled from a repository of the [University of Minho, Portugal](#). The dataset was created by Federico Soriano Palacios (2021), it integrates five different heart-related datasets combined over 11 common features that can be used to predict a possible heart disease. The five data sets are part of the (Janosi et al. 1989) that can be found in the UCI Machine Learning Repository that is originally sourced from the Hungarian Institute of Cardiology, the University Hospital of Zurich, the University Hospital of Basel, the V.A. Medical Center of Long Beach and Cleveland Clinic Foundation. Each row of the dataset contains 11 attributes that describe the patient’s age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG result, maximum heart rate achieved,

exercise induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, and the presence or absence of heart disease.

To build the classification model for the prediction of heart failure risk, an exploratory data analysis was conducted. Considering the distribution of the data, the features and their correlation to the target variable, it was decided to include all of them in the model with some standardization to improve results. Data was split into 70% for the training set and 30% for the test set. Different approaches were taken to evaluate which model would have the best accuracy in classification. We trained five different models, including Decision Tree, k-nearest neighbors (k-nn), Support Vector Machine (SVC) with an RBF kernel, Logistic Regression, and a Dummy Classifier (baseline). Each of the models were evaluated using a 5-fold cross-validation strategy. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The model with the highest performance scores across all metrics is the Logistic Regression which is the model that will be used for the rest of the project. The Python programming language (Van Rossum and Drake 2009) and the following Python packages were used to perform the analysis: NumPy (Harris and et al. 2020), pandas (McKinney 2010), Matplotlib (Hunter 2007), seaborn (Waskom 2021), Altair (VanderPlas and et al. 2018), scikit-learn (Pedregosa and et al. 2011), and requests (Reitz 2011).

Results & Discussion

We conducted some initial exploratory data analysis by observing the number of observations, data types of the features, and checked for missing values. We also inspected the categorical counts for some features to better understand their distribution. The distribution of the target variable was observed to be balanced with approximately equal representation of both classes.

To evaluate the usefulness of each feature to predict the heart disease, we investigated the distribution of each feature with respect to the target variable. We also visualized the correlation between features using a heatmap to identify any strong relationships.

Value Ranges for Data Validation

For the numerical features, reasonable value ranges were established to support data validation. Age typically falls between 20 and 90 years, while resting blood pressure generally ranges from 80 to 230 mm Hg. Cholesterol levels are commonly observed between 50 and 400 mm/dl; values near 50 may indicate malnutrition, whereas values above 400 are rare and may represent noise in the dataset. FastingBS is a binary indicator that takes the value 1 when fasting blood sugar exceeds 120 mg/dl and 0 otherwise. The MaxHR (maximum heart rate achieved) usually ranges from 60 to 202 beats per minute. Oldpeak, which reflects ST depression induced by exercise relative to rest, can take values from approximately -4 mm to $+6$ mm; negative values

indicate further ST depression during exercise, positive values indicate ST elevation, and zero indicates no change.

For the categorical features, Sex can be either F or M, while ChestPainType includes TA (typical angina), ATA (atypical angina), NAP (non-anginal pain), and ASY (asymptomatic). The RestingECG feature may take the values Normal, ST (indicating an ST-T wave abnormality), or LVH (suggesting left ventricular hypertrophy). ExerciseAngina is recorded as Y or N, and ST_Slope describes the slope of the peak-exercise ST segment, taking the values Up or Flat.

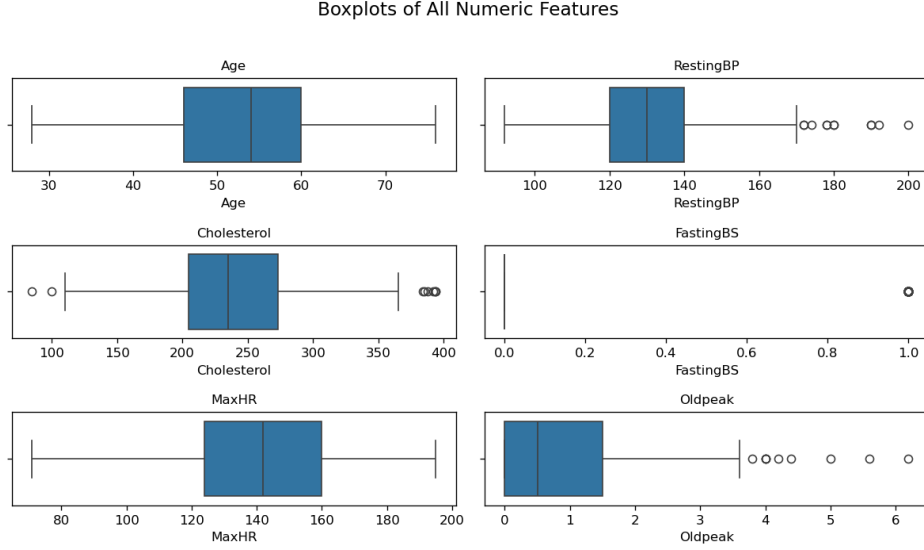


Figure 1: Boxplots of Numeric Features.

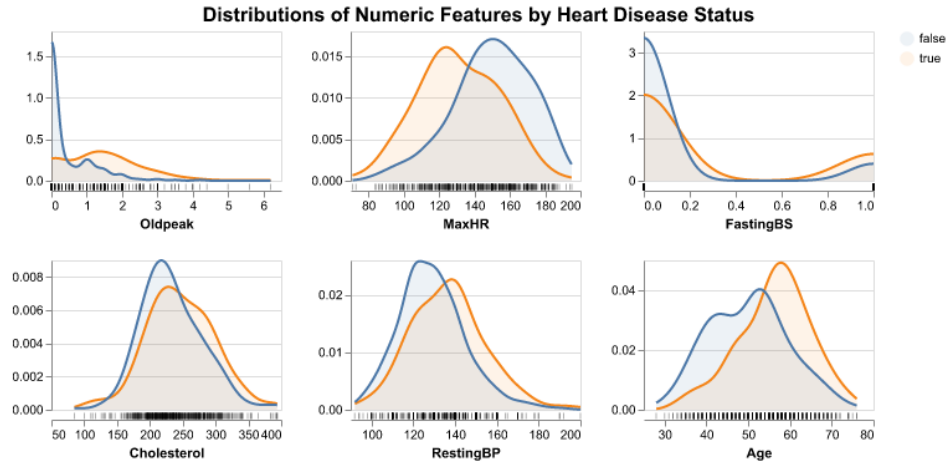


Figure 2: Distribution of categorical predictors, coloured by heart disease status.

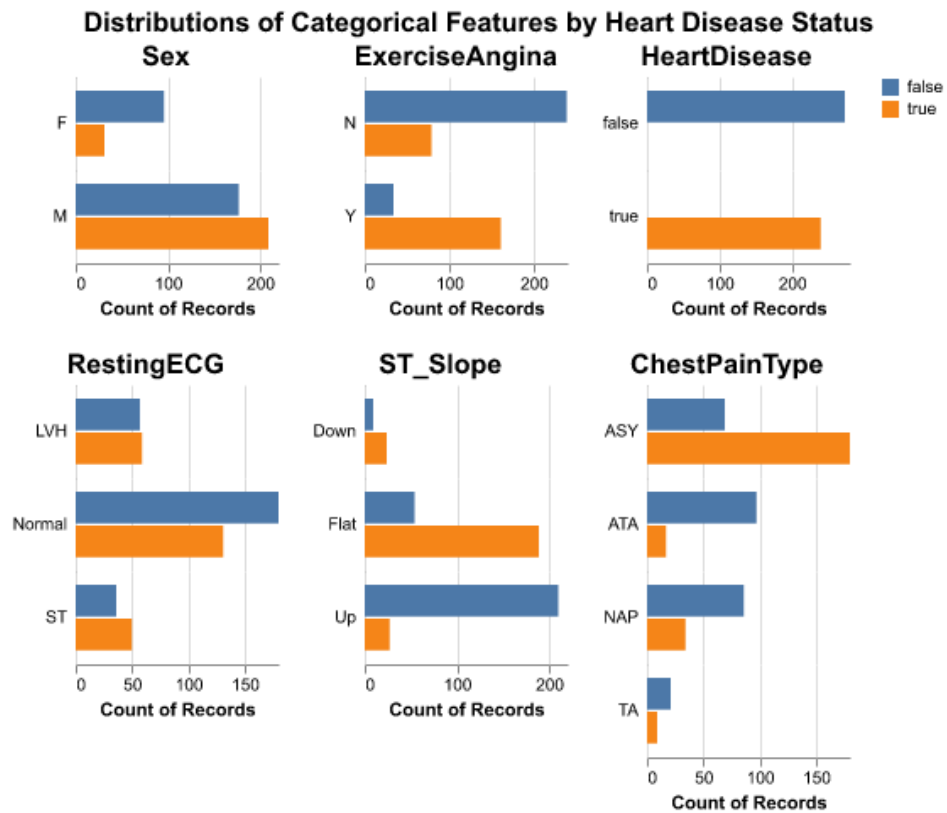


Figure 3: Comparison of numerical predictors, coloured by heart disease status.

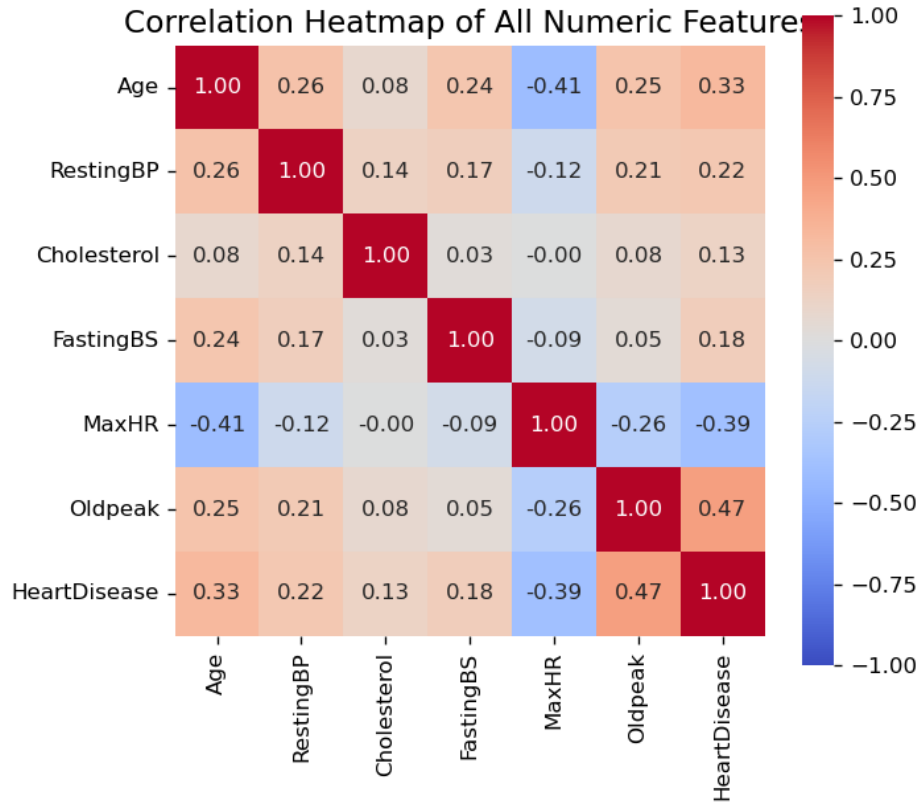


Figure 4: Correlation heatmap of all numerical features in the dataset.

We are not dropping any features as all features seem to be relevant to predicting heart disease based on the EDA.

Models Evaluation

We decided to initiate our analysis considering a set of classification models including Decision Tree, kNN, SVM (with RBF kernel), Logistic Regression, and a dummy classifier as a baseline. All models were trained and evaluated using a 5-fold cross-validation strategy. The performance classification metrics used to evaluate the models were accuracy, precision, recall and F1-score.

Initially, the models were trained using default hyperparameters to define a model that would be further optimized in subsequent steps for this project.

The configuration and results of the cross-validations are detailed below:

Table 1: Cross-validation 5-folds results for all models.

index	dummy	decision_tree	kNN	SVM	logistic_regression
fit_time	0.002 (+/- 0.000)	0.003 (+/- 0.000)	0.003 (+/- 0.000)	0.006 (+/- 0.002)	0.003 (+/- 0.000)
score_time	0.005 (+/- 0.000)	0.005 (+/- 0.000)	0.017 (+/- 0.008)	0.008 (+/- 0.001)	0.006 (+/- 0.001)
test_accuracy	0.532 (+/- 0.003)	0.760 (+/- 0.057)	0.819 (+/- 0.028)	0.854 (+/- 0.015)	0.850 (+/- 0.024)
train_accuracy	0.532 (+/- 0.001)	1.000 (+/- 0.000)	0.884 (+/- 0.006)	0.905 (+/- 0.002)	0.864 (+/- 0.006)
test_precision	0.000 (+/- 0.000)	0.745 (+/- 0.062)	0.804 (+/- 0.040)	0.837 (+/- 0.027)	0.833 (+/- 0.029)
train_precision	0.000 (+/- 0.000)	1.000 (+/- 0.000)	0.871 (+/- 0.009)	0.899 (+/- 0.008)	0.854 (+/- 0.003)
test_recall	0.000 (+/- 0.000)	0.742 (+/- 0.070)	0.812 (+/- 0.036)	0.854 (+/- 0.000)	0.850 (+/- 0.023)
train_recall	0.000 (+/- 0.000)	1.000 (+/- 0.000)	0.882 (+/- 0.009)	0.899 (+/- 0.009)	0.856 (+/- 0.014)
test_f1	0.000 (+/- 0.000)	0.743 (+/- 0.063)	0.808 (+/- 0.029)	0.846 (+/- 0.014)	0.841 (+/- 0.025)
train_f1	0.000 (+/- 0.000)	1.000 (+/- 0.000)	0.876 (+/- 0.007)	0.899 (+/- 0.002)	0.855 (+/- 0.007)

After performing the cross-validation, we observed that both the Logistic Regression and SVM models had the best performance across all metrics, with Logistic Regression slightly outperforming SVM all-around.

Based on this, we decided to choose the Logistic Regression model as our final model for predicting heart disease risk.

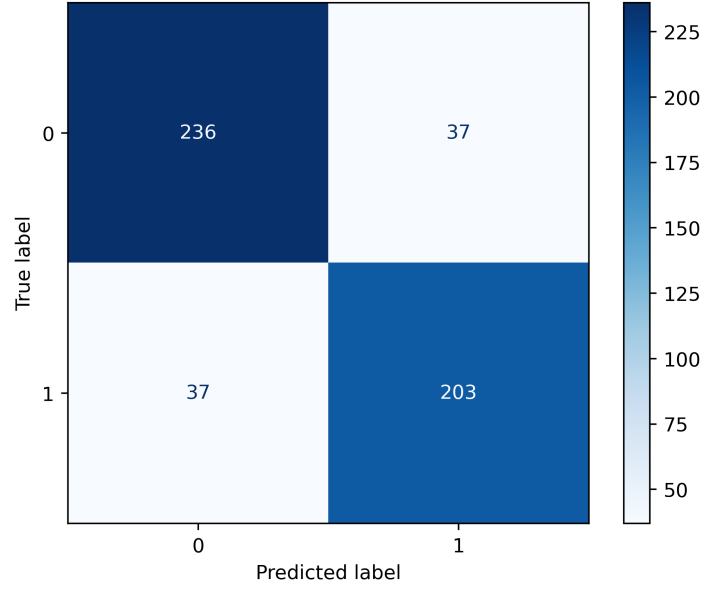


Figure 5: Confusion matrix for Logistic Regression model - Train set.

Evaluation on Test Set

The initial results of the Logistic Regression prediction model on the test set showed good performance across all metrics, as detailed in Table 2. This matched our expectations based on the cross-validation results, indicating that the model would generalize well on unseen data. This is already a positive outcome for a first iteration of the model and provides a solid foundation for further improvements. It also has the potential to be useful as a support tool for clinical screening of patients.

Table 2: Classification Report for Test Set.

	precision	recall	f1-score	support
0	0.869565	0.892857	0.881057	112
1	0.885714	0.861111	0.873239	108
accuracy	0.877273	0.877273	0.877273	0.877273
macro avg	0.87764	0.876984	0.877148	220
weighted avg	0.877493	0.877273	0.877219	220

Aside from the hyperparameter tuning that will be explored on future stages, this use case is well suited for Precision-Recall and Receiver Operating Characteristic (ROC) curve analysis to further evaluate the model's threshold settings and trade-offs between precision and recall.

This is a medical-related application, and it is crucial to minimize false negatives i.e., predicting a patient is not at risk when they actually are. These details are shown in the confusion matrix Figure 6 obtained from the test data set. Also, making the probability scores available for the predictions could increase the model’s utility, supporting the decision-making of healthcare professionals.

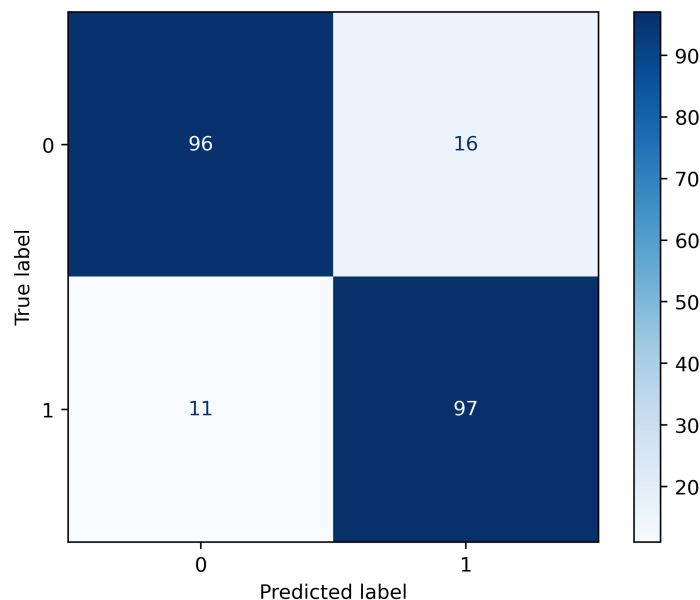


Figure 6: Confusion matrix for Logistic Regression model - Test set.

It is important to note that this project has certain limitations. The dataset used is relatively old and is based in Europe. There has been evidence that baseline heart-disease risk varies across racial and ethnic groups (Lewsey and Breathett 2021) which limits how well our results generalize to broader populations. Additionally, because we are relying on logistic regression, the model is not able to capture complex or non-linear relationships between the features and the target.

Finally, we could review the features coefficients from the Logistic Regression model to understand their influence on the predictions, enabling to better interpret the model’s decisions and confirm if we could drop any features in future iterations.

References

American Heart Association. 2025. “2025 Heart & Stroke Statistics Update Fact Sheet: Global Burden of Disease.” PDF. <https://professional.heart.org/-/media/phd-files->

- [2/science-news/2/2025-heart-and-stroke-stat-update/factsheets/2025-stats-update-fact-sheet-global-burden-of-disease.pdf](https://science-news/2/2025-heart-and-stroke-stat-update/factsheets/2025-stats-update-fact-sheet-global-burden-of-disease.pdf).
- Barnett, M. P., L. L. J. Koppes, and et al. 2020. “Cardiovascular Risk Factors: It’s Time to Focus on Variability!” *Frontiers in Cardiovascular Medicine* 7: Article 80. <https://doi.org/10.3389/fcvm.2020.00080>.
- Harris, C. R., and et al. 2020. “Array Programming with NumPy.” *Nature* 585: 357–62.
- Hunter, J. D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9 (3): 90–95.
- Janosi, A., W. Steinbrunn, M. Pfisterer, and R. Detrano. 1989. “Heart Disease.” UCI Machine Learning Repository; Dataset. <https://doi.org/10.24432/C52P4X>.
- Kokori, E., R. Patel, G. Olatunji, B. M. Ukoaka, I. C. Abraham, V. O. Ajekiigbe, J. M. Kwape, A. E. Babalola, N. G. Udam, and N. Aderinto. 2025. “Machine Learning in Predicting Heart Failure Survival: A Review of Current Models and Future Prospects.” *Heart Failure Reviews* 30 (2): 431–42. <https://doi.org/10.1007/s10741-024-10474-y>.
- Lewsey, S. C., and K. Breathett. 2021. “Racial and Ethnic Disparities in Heart Failure: Current State and Future Directions.”
- McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 51–56.
- Pedregosa, F., and et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30.
- Reitz, Kenneth. 2011. *Requests: HTTP for Humans*. <https://requests.readthedocs.io/en/master/>.
- Savarese, G., L. H. Lund, and P. M. Becher. 2023. “Global Burden of Heart Failure: A Comprehensive and Updated Review of Epidemiology.” *Cardiovascular Research* 118 (17): 3272–87. <https://doi.org/10.1093/cvr/cvac013>.
- Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
- VanderPlas, J., and et al. 2018. “Altair: Interactive Statistical Visualizations for Python.” *Journal of Open Source Software* 3 (32): 1057.
- Waskom, M. L. 2021. “Seaborn: Statistical Data Visualization.” *Journal of Open Source Software* 6 (60): 3021. <https://doi.org/10.21105/joss.03021>.
- World Health Organization. 2024. “Cardiovascular Diseases (CVDs) – Fact Sheet.” [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).