

Title: Using Logistic Regression to Predict the Presence of Heart Disease.

Name: Nzukiso Cele

Student Number: 2370259

Course Code: FAMH4004A, Health Analytics

Faculty: Health Sciences

Date of submission: 01 September 2025

**Table of Contents**

[Introduction 2](#_Toc876700615)

[Methods 3](#_Toc82545830)

[Data and Variable Description 3](#_Toc182146204)

[Data Pre-processing 3](#_Toc1570175508)

[Exploratory Data Analysis (EDA) 3](#_Toc698184800)

[Modelling Approach 3](#_Toc1575256628)

[Results 4](#_Toc825869814)

[Discussion 6](#_Toc417329303)

[Conclusion 7](#_Toc1720761949)

[References 7](#_Toc378336774)

## Introduction

Cardiovascular diseases (CVDs) continue to be the primary reason for mortalities worldwide. In 2019, they were linked to about 32% of mortalities worldwide, with ischemic heart disease causing 9.1 million fatalities and stroke contributing to 6.6 million (Di Cesare *et al.,* 2024 The aging global population is likely to drive a significant increase in the burden of CVDs over the coming decades (Chong *et al.,* 2024). Timely detection and precise diagnosis are crucial for lowering preventable mortality, yet predicting cardiac disease remains one of the most difficult challenges in clinical data analysis (Hassan *et al.,* 2022).

Machine learning (ML) tools have become valuable tools for supporting clinical decision-making by identifying hidden patterns in complex healthcare data (Rajkomar, Dean, and Kohane, 2019). Among these, logistic regression stands out for its interpretability and computational efficiency. Unlike more complex models, logistic regression provides transparency in how risk factors contribute to outcomes, which is crucial for clinical trust and adoption (Sidey-Gibbons and Sidey-Gibbons, 2019).

In this study, a logistic regression model was trained and tested to predict the occurrence of heart disease from patient records. Based on previous studies, we expected logistic regression to achieve moderate to high predictive accuracy (>70%) for detecting heart disease.

## Methods

### Data and Variable Description

The data used in this report was extracted from the Heart Attack Data Set spreadsheet.xlsx document. The dataset consists of 303 patients, with 14 variables describing demographic, clinical, and test features. The outcome variable is target (1 = presence of heart disease, 0 = absence).

Predictor variables included: demographics (age, sex), clinical measures (“chest pain type (*cp*), resting blood pressure (*trestbps*), serum cholesterol (*chol*), fasting blood sugar (*fbs*), maximum heart rate achieved (*thalach*), exercise-induced angina (*exang*), ST depression (*oldpeak*), slope of the ST segment (*slope*), number of major vessels visualised (*ca*)) and diagnostic tests (resting ECG results (*restecg*), thalassemia classification (thal)”).

### Data Pre-processing

The raw dataset required cleaning and formatting. Data originally presented in a single comma-separated column was split into structured variables with aligned headers. Each variable was checked for missing values; none were identified. Categorical variables (*cp, restecg, thal, slope, ca*) were encoded numerically as provided.

### Exploratory Data Analysis (EDA)

EDA combined descriptive statistics with visualisations to assess variable distributions and potential predictors of heart disease. Key analyses included correlation matrix (heatmap) to identify linear associations among predictors and with the target, histograms for age and cholesterol distributions and bar plots comparing prevalence of heart disease by gender and resting ECG category. These steps guided expectations for model performance and highlighted clinically relevant features.

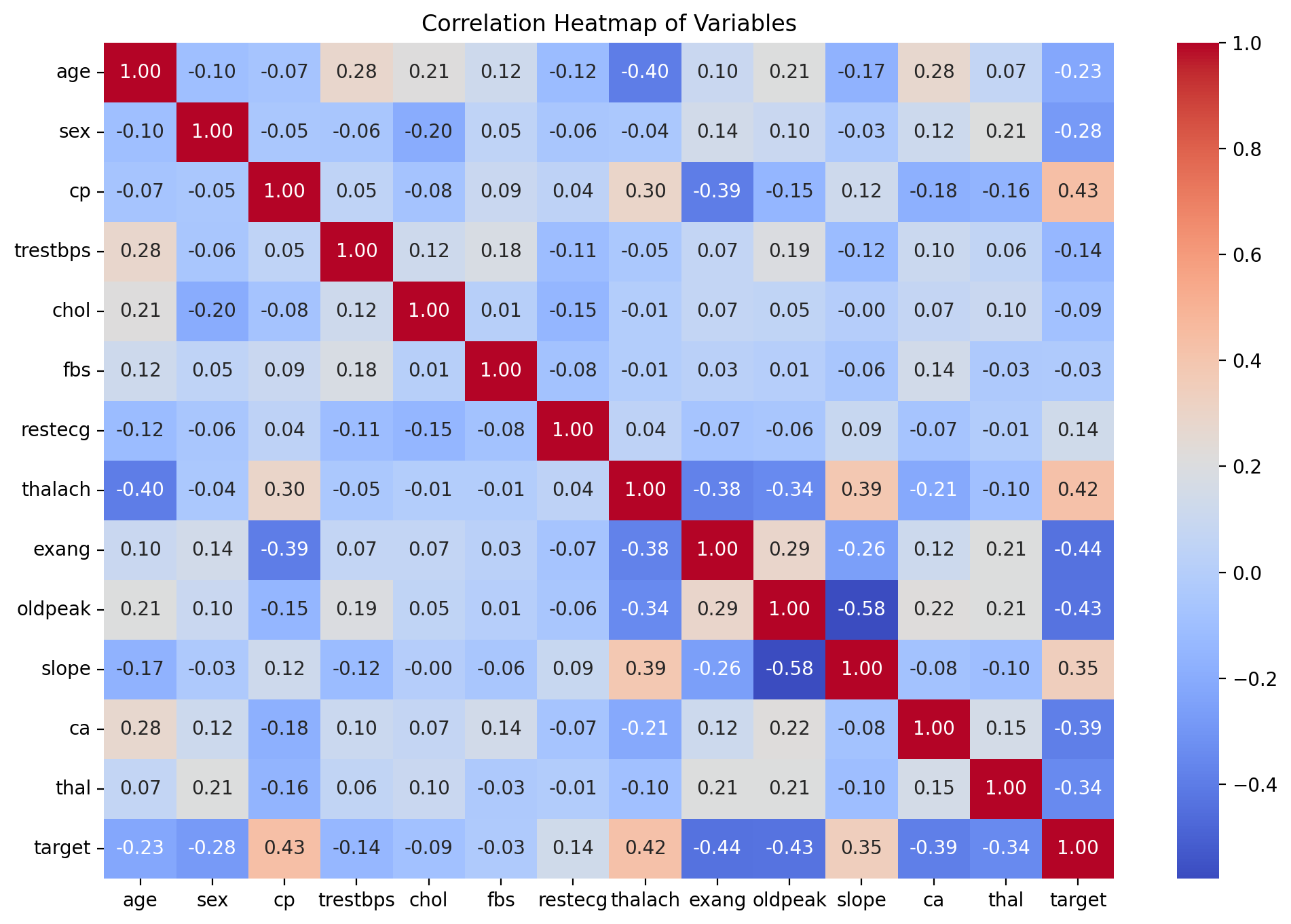
### Modelling Approach

A logistic regression classifier was trained using scikit-learn. The dataset was split into training (80%) and testing (20%) subsets.

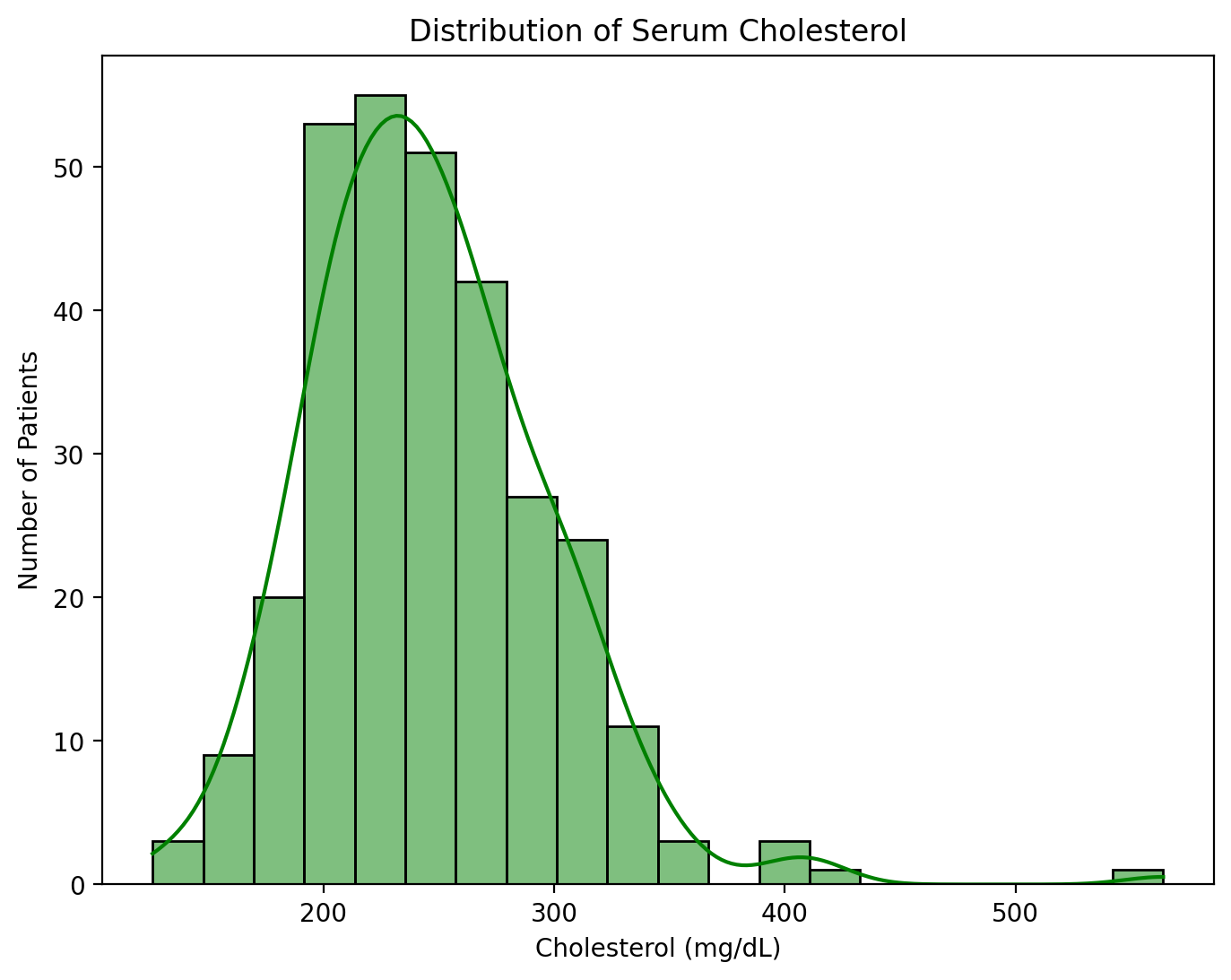
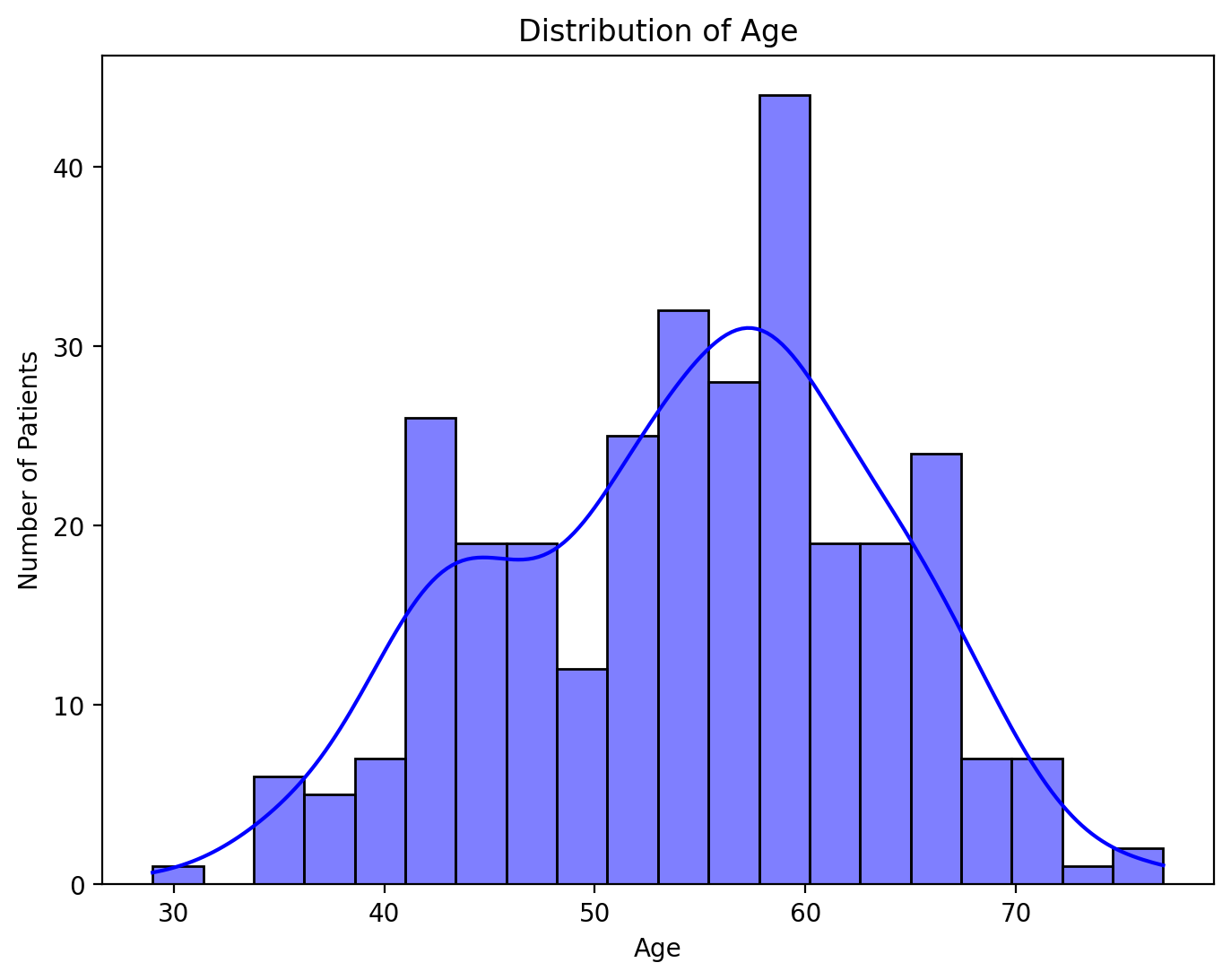
Model performance was evaluated using accuracy (overall correctness of predictions), confusion matrix (visual representation of classification performance), precision (proportion of predicted positives that were correct), recall (ability to identify true positives), and F1-score (balance between precision and recall). This ensured a comprehensive understanding of the model’s predictive capability.

## Results

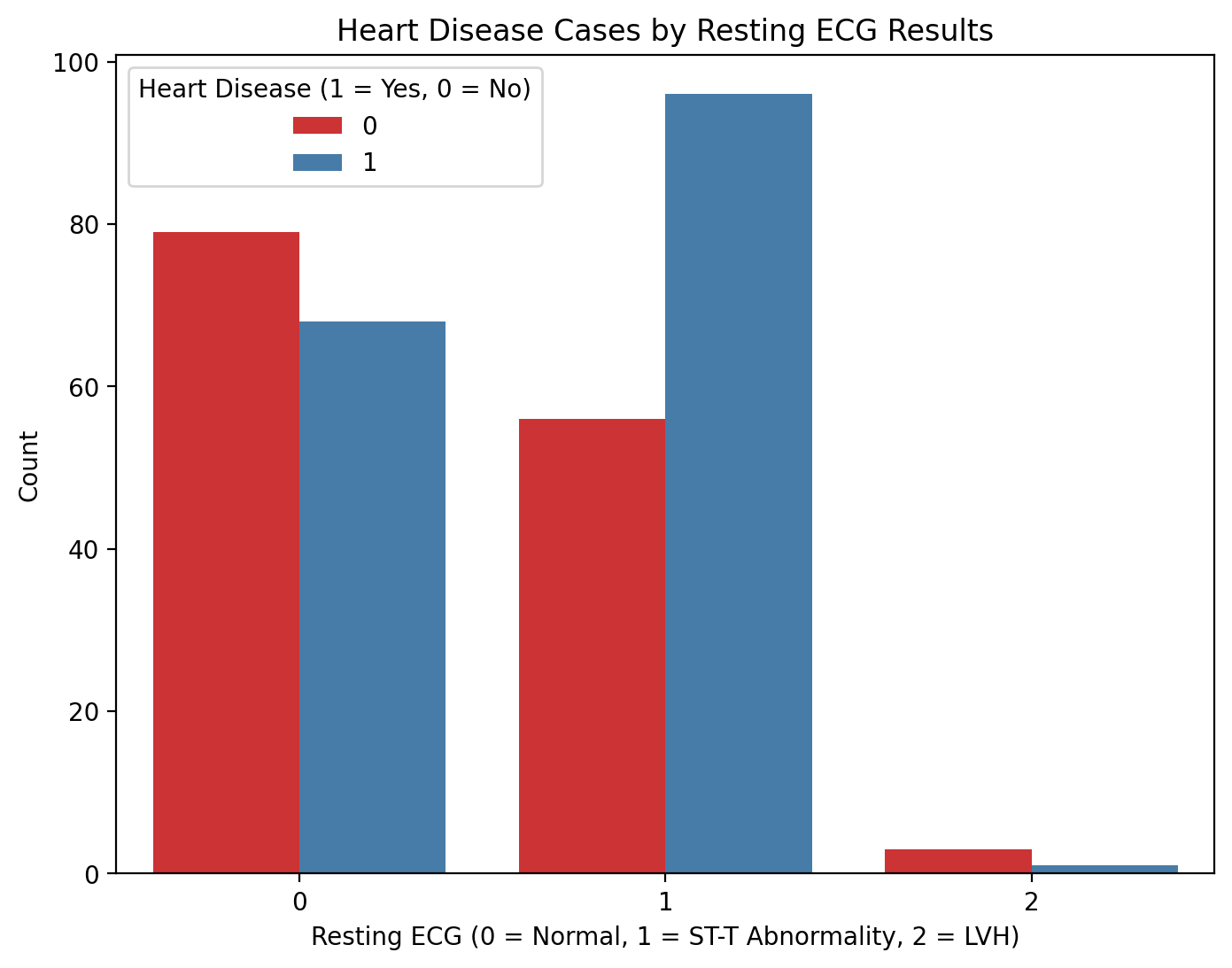
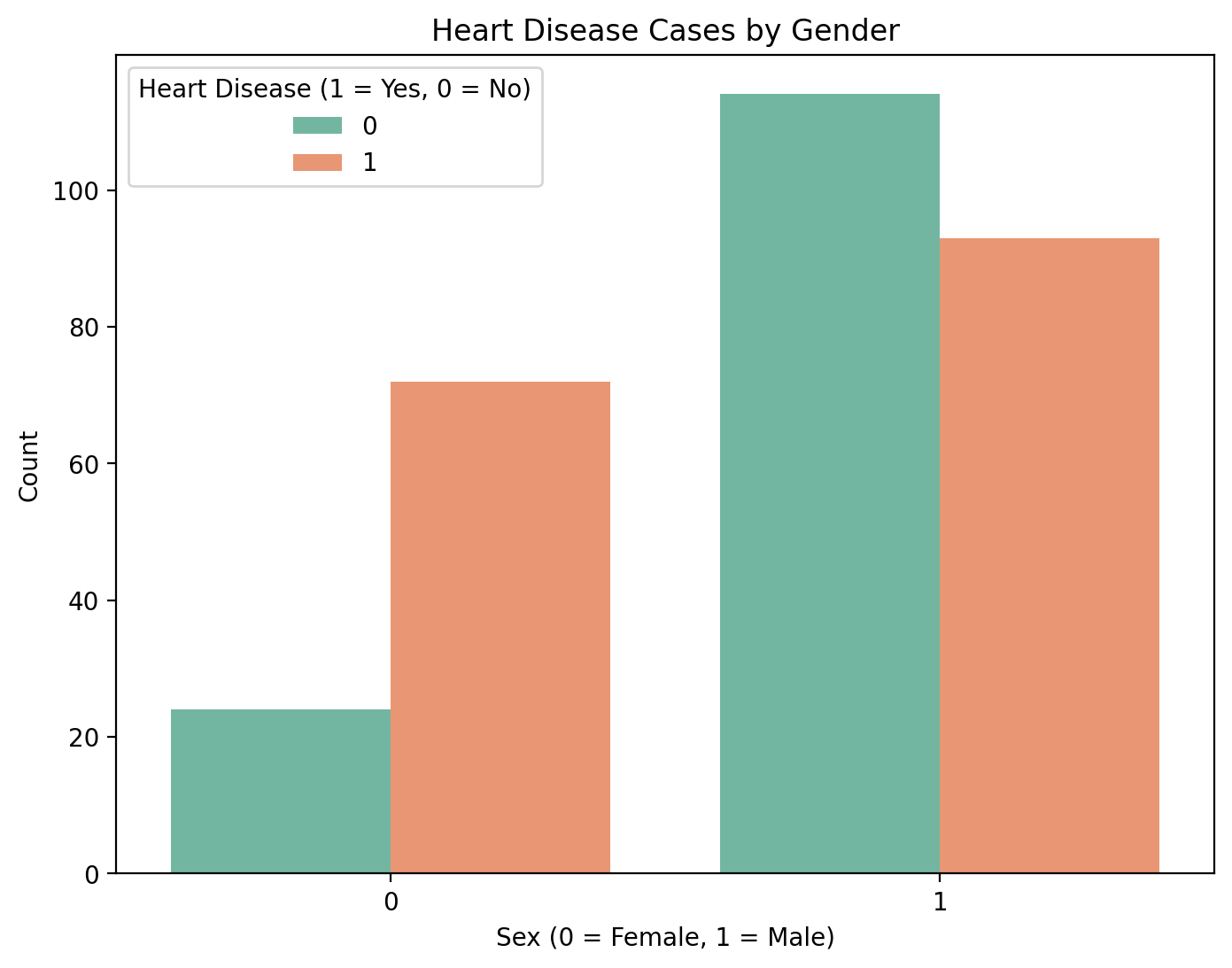
To evaluate the ability of logistic regression to predict heart disease, the results are organised into two parts: exploratory analysis of the dataset and performance of the trained model.



**Figure 1**: Correlation heatmap of the dataset. Chest pain type (cp) showed the strongest positive correlation with the target variable. Thalach and slope demonstrated moderate positive correlations, while restecg exhibited a weak correlation. Most other variables displayed weak or negative correlations with the target.



**Figure 2:** Distribution of age (left, blue) and cholesterol levels (right, green). The age distribution revealed that most patients were between 50 and 60 years old. Cholesterol levels were predominantly concentrated between 200 and 300 mg/dL, with a small number of patients presenting values exceeding 500 mg/dL.



**Figure 3**: Distribution of heart disease cases by gender (left) and resting electrocardiogram (ECG) results (right). The gender distribution indicates that although the dataset contains a larger number of male patients overall, females exhibited a higher proportion of positive heart disease cases. The ECG distribution shows that patients with ST–T abnormalities had the highest prevalence of heart disease, followed by those with left ventricular hypertrophy, while patients with normal ECGs had the lowest prevalence.

**Table 1**: Confusion matrix for the model, indicating the classification distribution. Out of 91 test cases, the model correctly classified 69 instances, resulting in an accuracy of 75.8%.

|  |  |  |
| --- | --- | --- |
|  | Predicted: No Disease | Predicted: Disease |
| Actual: No Disease | 28 | 13 |
| Actual: Disease | 9 | 41 |

The classification metrics are summarised in Table 2 below.

**Table 2**: Classification report. The model achieved high sensitivity (82%). However, specificity was lower (68%).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.757 | 0.683 | 0.718 | 41 |
| 1 | 0.759 | 0.820 | 0.788 | 50 |
| accuracy |  |  | 0.758 | 91 |
| Macro avg | 0.758 | 0.751 | 0.753 | 91 |
| Weighted avg | 0.758 | 0.758 | 0.757 | 91 |

## Discussion

The results demonstrate that logistic regression was able to predict heart disease with an accuracy of approximately 76%. Importantly, the model achieved a recall (sensitivity) of 82%, meaning it correctly identified most patients with heart disease. This is desirable in a clinical context because it minimises the number of missed diagnoses (false negatives). However, specificity was lower (68%), reflecting the model’s tendency to misclassify some healthy individuals as having heart disease. This trade-off is common in diagnostic prediction, where sensitivity is often prioritised over specificity to ensure that true cases are not overlooked (Rajkomar, Dean, and Kohane, 2019).

The exploratory analysis showed that chest pain type and exercise-related measures (thalach, slope, oldpeak) were moderately associated with heart disease, while cholesterol and resting blood pressure had weaker associations. These findings align with prior studies, which emphasise chest pain and exercise tolerance as key indicators of ischemic heart disease (Liu *et al.,* 2022).

Gender distributions in the dataset highlighted that although men were more numerous, women had a higher proportion of positive diagnoses. This reflects established evidence that cardiovascular disease often presents more severely in women, and sometimes with atypical symptoms (Mehta *et al.,* 2016). Similarly, the ECG distribution confirmed that ST-T abnormalities are an important diagnostic marker, consistent with clinical guidelines (Arnett *et al.,* 2019).

Nevertheless, limitations such as moderate specificity and relatively small dataset size suggest the need for further optimisation and validation on larger, more diverse datasets.

## Conclusion

This study aimed to build and evaluate a logistic regression model to predict heart disease. The model achieved 75.8% accuracy, with high sensitivity (82%) but moderate specificity (68%). The results indicate that logistic regression is effective at detecting most cases of heart disease, though it produces some false positives.

Exploratory data analysis confirmed that chest pain type, exercise tolerance, and ECG abnormalities are moderately associated with heart disease.

## References

Arnett, D.K., Blumenthal, R.S., Albert, M.A., Buroker, A.B., Goldberger, Z.D., Hahn, E.J., Himmelfarb, C.D., Khera, A., Lloyd-Jones, D., McEvoy, J.W. and Michos, E.D., 2019. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology*, *74*(10), pp.1376-1414. Available at: [2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines | JACC](https://www.jacc.org/doi/abs/10.1016/j.jacc.2019.03.009) (Acceessd: 31 August 2025)

Chong, B., Jayabaskaran, J., Jauhari, S.M., Chan, S.P., Goh, R., Kueh, M.T.W., Li, H., Chin, Y.H., Kong, G., Anand, V.V. and Wang, J.W., 2024. Global burden of cardiovascular diseases: projections from 2025 to 2050. *European journal of preventive cardiology*, p.zwae281. Available at: [Global burden of cardiovascular diseases: projections from 2025 to 2050 - PubMed](https://pubmed.ncbi.nlm.nih.gov/39270739/) (Accessed: 28 August 2025)

Di Cesare, M., Perel, P., Taylor, S., Kabudula, C., Bixby, H., Gaziano, T.A., McGhie, D.V., Mwangi, J., Pervan, B., Narula, J. and Pineiro, D., 2024. The heart of the world. *Global heart*, *19*(1), p.11. Available at: [The Heart of the World - PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC10809869/) (Accessed: 28 August 2025)

Hassan, C.A.U., Iqbal, J., Irfan, R., Hussain, S., Algarni, A.D., Bukhari, S.S.H., Alturki, N. and Ullah, S.S., 2022. Effectively predicting the presence of coronary heart disease using machine learning classifiers. *Sensors*, *22*(19), p.7227. Available at: [Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers - PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC9573101/) (Accessed: 28 August 2025)

Liu, W., Liu, X., Liu, T., Xie, Y., He, X., Zuo, H. and Zeng, H., 2022. The value of cardiopulmonary exercise testing in predicting the severity of coronary artery disease. *Journal of Clinical Medicine*, *11*(14), p.4170. Available at: [The Value of Cardiopulmonary Exercise Testing in Predicting the Severity of Coronary Artery Disease](https://www.mdpi.com/2077-0383/11/14/4170) (Accessed: 31 August 2025)

Mehta, L.S., Beckie, T.M., DeVon, H.A., Grines, C.L., Krumholz, H.M., Johnson, M.N., Lindley, K.J., Vaccarino, V., Wang, T.Y., Watson, K.E. and Wenger, N.K., 2016. Acute myocardial infarction in women: a scientific statement from the American Heart Association. *Circulation*, *133*(9), pp.916-947. Available at: [Acute Myocardial Infarction in Women | Circulation](https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000000351) (Accessed: 31 August 2025)

Rajkomar, A., Dean, J. and Kohane, I., 2019. Machine learning in medicine. *New England Journal of Medicine*, *380*(14), pp.1347-1358. Available at: [Machine Learning in Medicine | New England Journal of Medicine](https://www.nejm.org/doi/full/10.1056/NEJMra1814259?casa_token=I6f1cMuc-tAAAAAA%3Ap8OzRV4DkG7Vld-MP5dFiccwde4_oWye7Hq_vjzJQqEX8xtGdWeTH0TwVpRB4wmJfl9EKlSf6f1GkJg) (Accessed: 28 August 2025)

Sidey-Gibbons, J.A. and Sidey-Gibbons, C.J., 2019. Machine learning in medicine: a practical introduction. *BMC medical research methodology*, *19*(1), p.64. Available at: [Machine learning in medicine: a practical introduction | BMC Medical Research Methodology](https://link.springer.com/article/10.1186/s12874-019-0681-4) (Accessed: 28 August 2025)