Logistic Regression

1 September 2025

**Table of Contents**

1. Introduction………………………………………….1
2. Aim…………………………………………………..1
3. Method
   1. Data description……………………………….…1
   2. Tools…………………………………………..….2
   3. Model justification……………………….............2
   4. Procedural steps for analysis……………………..2
4. Results
   1. Correlation analysis…………………………...….2
   2. Age distribution…………………………………..3
   3. Cholesterol distribution…………………………..3
   4. Heart disease cases by gender distribution……….3
   5. Heart disease cases by resting ECG distribution…4
   6. Model Evaluation………………………………....4
5. Discussion………………………………….………....5
6. Conclusion…………………………………………….5

**1. Introduction**

According to Christodoulou et al. (2019), cardiovascular illnesses continue to be the world's leading cause of morbidity and mortality, highlighting the significance of early detection and prevention. Heart failure, stroke, hypertension, coronary artery disease, and other conditions affecting the heart and blood vessels are collectively called cardiovascular diseases (CVDs). It frequently originates from a confluence of environmental, behavioural, and genetic factors, making it one of the major causes of death worldwide.

Clinical datasets and predictive analytics are a useful means of assisting in medical decision-making. A statistical classification technique called logistic regression is frequently used in the healthcare industry because it produces findings that are easy to understand and can compete with more sophisticated machine learning algorithms (James et al., 2021; Christodoulou et al., 2019). According to Kumar et al. (2022), classification it is the process of classifying observations into predetermined groups according to their characteristics. When the goal variable is binary, like predicting if a patient will have a heart attack (yes/no), logistic regression is one of the most used classification methods.

This report utilizes logistic regression on a heart attack dataset to estimate the probability of a heart attack based on patient features. This illustrates the importance of classification in the medical field and underscores the factors that must be taken into account when choosing algorithms for these sensitive applications, especially the balance between accuracy and interpretability.

**2. Aim of this report**

* To describe the function of categorization in the medical field, emphasizing its significance for clinical decision-making, risk assessment, and illness detection.
* To illustrate how logistic regression can be used as a classification technique to forecast binary health outcomes.
* To examine, taking into account aspects like accuracy, transparency, and clinical usability, if logistic regression is more appropriate for use in healthcare applications than more intricate machine learning models.

**3. Methodology**

**3.1 Dataset description**

* **Numeric variables:** age (years), tresbps (resting blood pressure (mm Hg)), chol (serum cholesterol (mg/dl)), thalach (maximum heart rate achieved), oldpeak (ST depression induced by exercise), etc.
* **Categorical variables:** sex (male/female), cp (chest pain type), fbs (fasting blood sugar > 120 mg/dl), restecg (resting electrocardiographic results), exang (exercise-induced angina), slope (slope of the peak exercise ST segment), ca (number of major vessels colored by fluoroscopy), thal (thalassemia (normal/fixed defect/reversible defect)) and target (1 = Disease, 0 = No disease).
  1. **Tools**
* Pandas and NumPy- used for data loading and processing.
* Scikit-learn- used for regression modelling and assessment.
* Matplotlib and Seaborn- used for data visualization.
  1. **Model justification**

Because the dataset consists of a binary outcome (heart attack: yes/no), logistic regression was selected as the method of choice for classification. Its simplicity, interpretability, and capacity to calculate risk probabilities are all critical for clinical decision-making which makes it a popular tool in the healthcare industry (Hosmer et al., 2013; Christodoulou et al., 2019).

* 1. **Procedural steps for Analysis**
     1. Data loading and Preprocessing

-Heart attack dataset(csv) was loaded into Python

* + 1. Exploratory data analysis (EDA)

-Correlation analysis looked at how variables are related to one another.

-The distribution of age and cholesterol level was illustrated.

-The distribution of heart attack cases by gender was presented.

-The distribution of heart attack cases by resting ECG results was presented.

* + 1. Model Training

-Test (20%) and training (80%) sets of the dataset were separated.

-The training data was fitted to a logistic regression model.

* + 1. Model evaluation

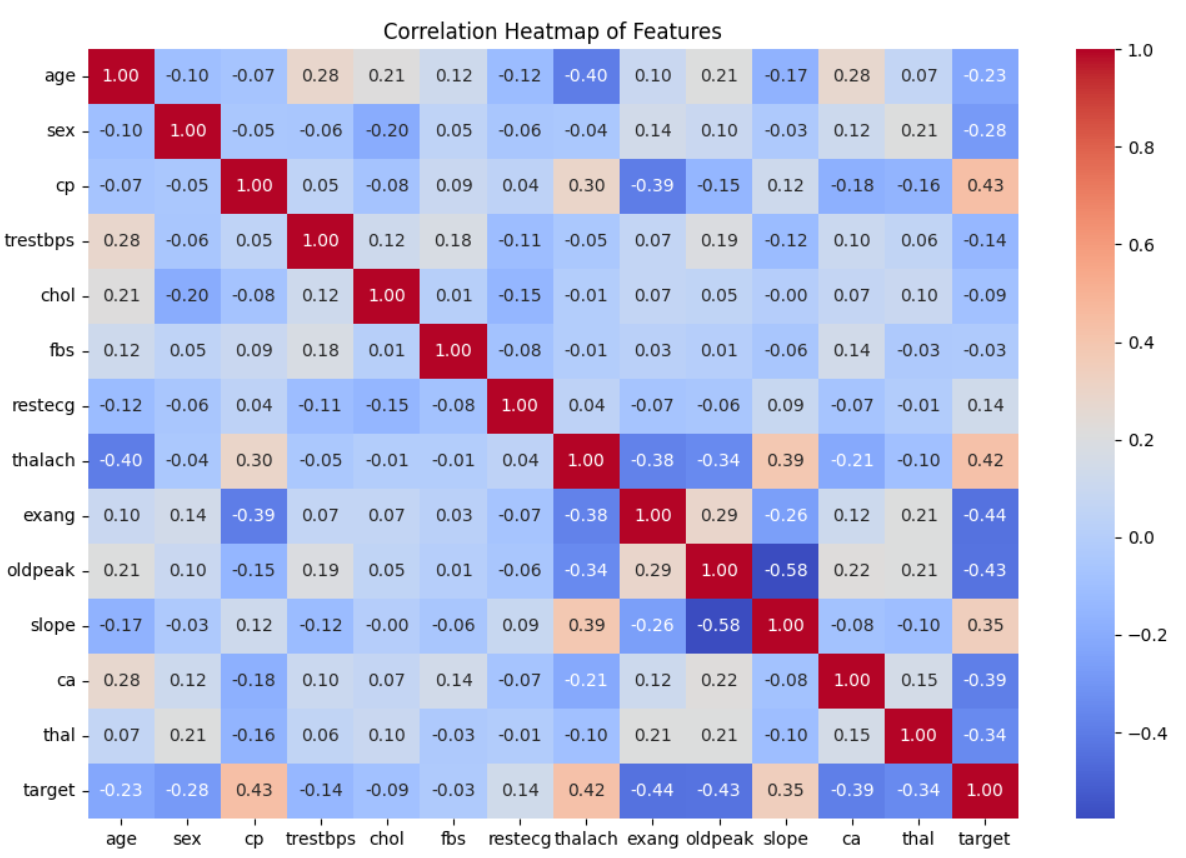
-The model's F1-score, recall, accuracy, and precision were assessed.

-To evaluate categorization performance, a confusion matrix was created.

-Predictive power and clinical relevance were used to interpret the results.

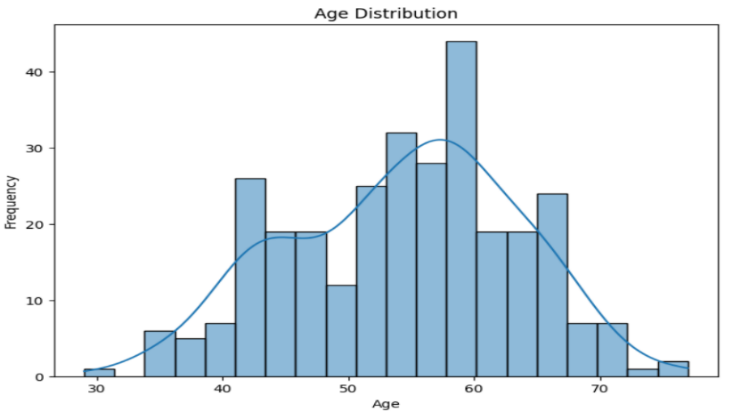
**4. Results**

**4.1 Figure 1. Correlation analysis**



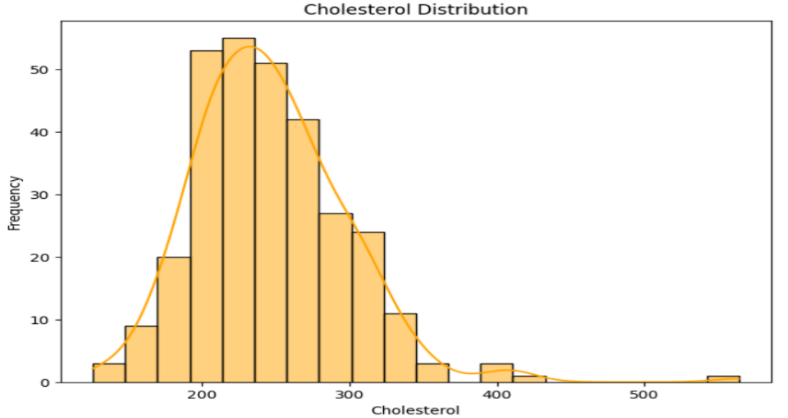
Description: The assumptions of logistic regression were acceptable because the predictors did not exhibit extreme multicollinearity (no very high |r| pairs). While inter-feature correlations were smaller, several features showed moderate connections with the target, indicating that each variable adds complementing information.

**4.2 Figure 2. Age distribution**



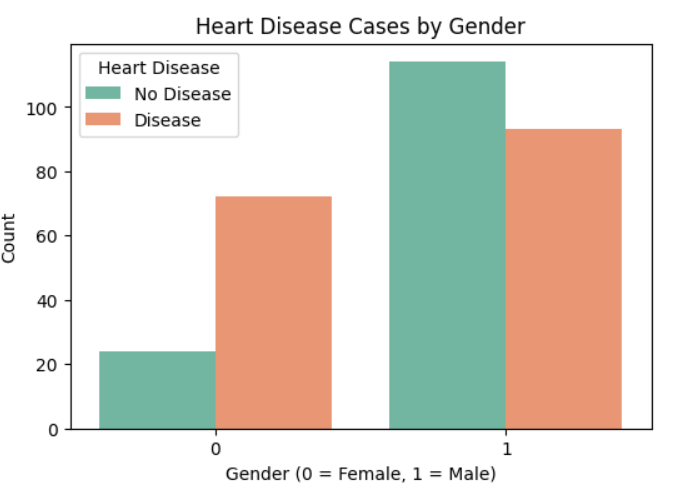
Description: With fewer extremely young or very old patients, ages were concentrated around mid-to-late adulthood; the shape was moderately skewed but devoid of extreme outliers.

**4.3Figure 3. Cholesterol level distribution**



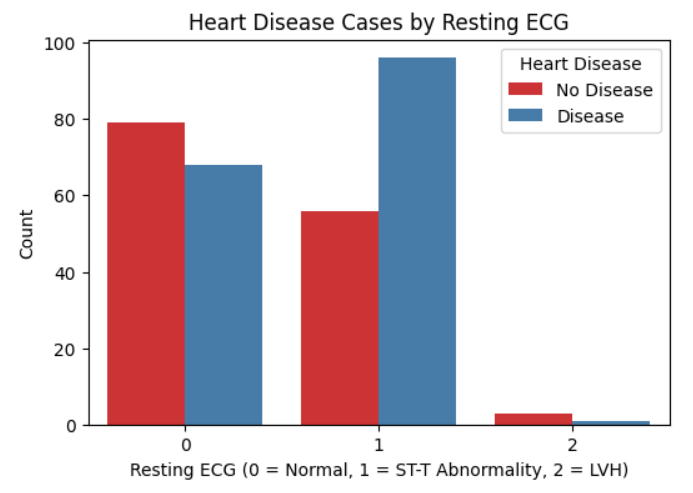
Description: There are a few higher-risk outliers among the cholesterol readings, but the majority of observations are in the middle of the range with a right tail.

**4.4 Figure 4.** **Distribution of heart attack cases by gender**



Description: Males were more prevalent than females in the dataset, and males in this group had a larger percentage of disease.

**4.5 Figure 5.** **Distribution of heart disease cases by resting ECG results**



Description: As clinically expected, the number of cases with ST-T anomalies was higher in the sick group than in the normal ECG group.

**4.6 Model evaluation**

Table 1. Metrics

|  |  |
| --- | --- |
| Accuracy | 0.8524590163934426 |
| Precision | 0.8709677419354839 |
| Recall | 0.84375 |
| F1-score | 0.8571428571428571 |

Figure 6. Confusion matrix

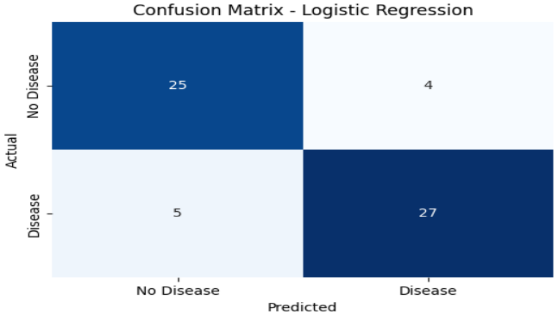
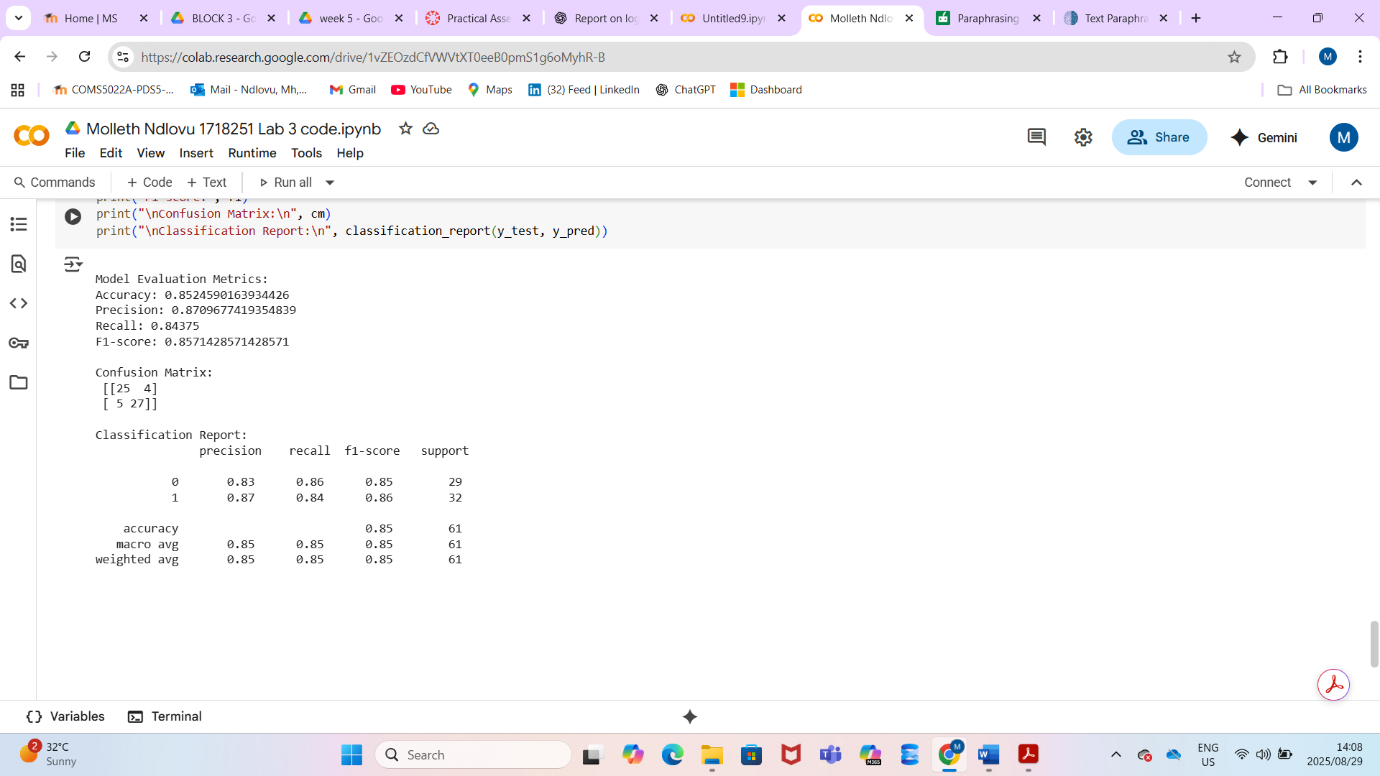


Figure 7. Classification report



Description: In figure 6 & 7 and table 1, Performance demonstrates a well-balanced classifier with a low false-positive rate (precision) and good positive-case identification (recall). When a model's accuracy exceeds 75%, it can be deemed successful.

**5. Discussion**

According to the correlation heatmap, there were moderate associations between heart disease and age, cholesterol, and maximum heart rate. Positive correlations indicated that risk increased with age and cholesterol levels, while negative correlations indicated a protective effect. According to the age and cholesterol distribution plots, patients with heart disease tended to be older and had higher cholesterol levels than those without. Additionally, the resting ECG chart showed that abnormal ECG results were more commonly linked to positive heart disease cases, while the gender-based graph showed a higher prevalence of heart disease among males. All of these results point to a significant interaction between diagnostic, clinical, and demographic factors in predicting cardiovascular risk, which supports their inclusion in predictive modelling.

The findings show that logistic regression can be a useful method for estimating a patient's risk of developing heart disease as the majority of instances were accurately identified with 85.2% accuracy rate. While the recall (84.4%) indicates that the model was successful in identifying the majority of actual positive cases, the precision (87.1%) indicates that the majority of patients projected to have heart disease were in fact true cases. The model's consistent performance across both measures is confirmed by the F1-score (85.1%), which strikes a balance between precision and recall.

The interpretability of logistic regression is one of its main advantages in the medical field. The link between risk factors (such age, cholesterol, and ECG readings) and disease outcomes can be easily examined using logistic regression, in contrast to more intricate black-box models (Sharma & Rani, 2021). Logistic regression does have several drawbacks, though. It makes the assumption that predictors and result log-odds have a linear relationship, which might not accurately represent intricate biomedical interactions (Sidey-Gibbons & Sidey-Gibbons, 2019). nonetheless, it is appropriate for structured healthcare data, like the heart disease dataset utilized here, because of its resilience and powerful performance on moderately big datasets.

The evaluation criteria collectively indicate that logistic regression is a useful tool for clinically relevant performance in grouping patients into illness/no disease categories. These findings align with other research that supports the use of logistic regression as a baseline model in predictive analytics for medical care (Jain et al., 2021; Patel et al., 2022).

**6. Conclusion**

In conclusion, This study proved that using patient clinical information, logistic regression is a good and efficient way to predict heart disease. Reliability in patient classification was demonstrated by the model's balanced accuracy, precision, recall, and F1-score. It is useful for assisting medical decision-making because of its interpretability, which further supports its function as a useful tool in healthcare analytics.

**References**

Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y. and Van Calster, B. (2019) ‘A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models’, Journal of Clinical Epidemiology, 110, pp. 12–22. doi:10.1016/j.jclinepi.2019.02.004.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied Logistic Regression. 3rd ed. Hoboken: Wiley.

Jain, P., Sharma, V. & Choudhary, A., 2021. Application of logistic regression in healthcare prediction models. Journal of Healthcare Informatics Research, 5(3), pp.245-259.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) An Introduction to Statistical Learning with Applications in R and Python. 2nd edn. New York: Springer. doi:10.1007/978-1-0716-1418-1.

Kumar, V., Minz, S. and Kumar, S., 2022. Classification techniques in healthcare applications: A review and analysis. Health Information Science and Systems, 10(1), pp.1–16.

Patel, S., Kumar, A. & Singh, R., 2022. Machine learning applications in healthcare: A focus on logistic regression. Health Information Science and Systems, 10(1), pp.1-10.

Sidey-Gibbons, J. & Sidey-Gibbons, C., 2019. Machine learning in medicine: A practical introduction. BMC Medical Research Methodology, 19(1), p.64.

Sharma, S. & Rani, R., 2021. Explainable artificial intelligence models in healthcare: A review. Computer Methods and Programs in Biomedicine, 200, 105886.