Journal of Media Horizons
ISSN: 2710-4060 2710-4052 Volume 6, Issue 7, 2025

LOGISTIC REGRESSION MODELING OF DISEASE RISK FACTORS: A CASE STUDY

Dr Arzoo Kanwal[*1], Qurratulain Hanif[2]

[1]*Associate Professor in Statistics, Institute of Numerical Science Gomal University D.I.Khan,*
[2]*Biostatistician, Avicenna Hospital*

[1]arzookanwal786786@gmail.com, [2]anieebhatti9@gmail.com

DOI: https://doi.org/10.5281/zenodo.17876269

Keywords
Logistic Regression, Cardiovascular Disease, Risk Prediction, Clinical Diagnostics,

Article History
Received: 14 October 2025
Accepted: 24 November 2025
Published: 10 December 2025

Corresponding Author: *
Dr Arzoo Kanwal

Abstract
Cardiovascular disease remains a major global health challenge, requiring accurate and interpretable predictive tools for early diagnosis and effective prevention. This study develops a multivariable logistic regression model to identify significant demographic and clinical factors associated with heart disease. Using a structured clinical dataset, the analysis examined key predictors including age, chest pain type, resting blood pressure, cholesterol, exercise-induced angina, and ST depression. The results reveal that age, abnormal chest pain categories, exercise-induced angina, and ST depression are strong independent predictors of disease presence. The final model demonstrated excellent performance with high classification accuracy and an AUC exceeding 0.80, indicating strong discriminative ability. Model diagnostics confirmed adequate goodness-of-fit and stable coefficient estimates. These findings highlight the importance of integrating clinical, physiological, and stress-test variables to enhance cardiovascular risk assessment. The study reinforces the value of logistic regression as an interpretable and reliable approach for informing clinical decision-making and provides a foundation for future research incorporating larger samples, additional biomarkers, and external validation to improve predictive precision.

INTRODUCTION

Cardiovascular diseases (CVDs) continue to be among the leading causes of morbidity and mortality worldwide, posing a major public health burden in both developed and developing countries. Early identification of individuals at high risk is therefore essential for timely intervention, prevention, and evidence-based clinical decision-making. Statistical modeling plays a central role in risk prediction, particularly logistic regression, which offers interpretability, robustness, and the ability to estimate independent associations between risk factors and disease outcomes. Logistic regression is widely applied in medical decision support systems, epidemiological surveys, and diagnostic research because it allows simultaneous examination of multiple predictors such as age, blood pressure, cholesterol level, exercise patterns, and demographic variables while controlling for confounding and assessing effect sizes through odds ratios. In recent years, with increasing availability of health datasets, researchers have combined classical statistical models with machine-learning strategies to improve prediction performance. While advanced models such as random forests, SVMs, and boosting algorithms often provide higher predictive accuracy, logistic regression remains

the preferred choice in clinical environments due to its transparency and interpretability. For populations where resources are limited, interpretable models help clinicians understand how risk factors influence disease, enhancing trust and adoption. The present study utilizes a clinical dataset to develop a multivariable logistic regression model for predicting disease presence based on key demographic and medical variables. The study evaluates statistical significance, predictive accuracy, and diagnostic measures to offer a comprehensive understanding of the factors contributing to disease risk. By integrating clinical reasoning with empirical modeling, this research contributes to improved prediction frameworks and informs future analytical approaches.

Numerous studies have explored the use of logistic regression in medical diagnosis and risk prediction. Hosmer and Lemeshow (2000) established logistic regression as a fundamental tool for modeling binary clinical outcomes. Harrell (2015) emphasized the importance of model diagnostics, calibration, and validation in developing reliable prediction models. Clinical prediction frameworks were further advanced by Steyerberg (2009), who detailed methods for model development, performance assessment, and updating. The Framingham Heart Study remains one of the most influential sources for cardiovascular risk modeling. Wilson et al. (1998) and D'Agostino et al. (2008) developed multivariable risk functions based on age, blood pressure, cholesterol, diabetes, and smoking, forming the benchmark for global CVD prediction. Artigao-Ródenas et al. (2013) validated these models in non-U.S. populations and noted the need for regional recalibration. Comparative studies have demonstrated varying results regarding predictive performance. Bharti et al. (2021) compared logistic regression with random forest and SVM classifiers, finding that while machine-learning methods achieved slightly higher accuracy, logistic regression remained more interpretable. Alfadli (2022) and Li (2025) similarly found that logistic regression performed well when appropriate feature selection and data preprocessing were applied. Yaqoob (2023) provided a comprehensive review showing that classical statistical models still outperform complex algorithms when datasets are small or moderately sized. From a methodological perspective, Peduzzi et al. (1996) and Riley et al. (2019) emphasized appropriate sample size, event-per-variable ratios, and avoidance of overfitting. Collins et al. (2015) introduced the TRIPOD guidelines, promoting transparent reporting for prediction models. To enhance model performance, researchers have used regularization strategies such as the Elastic Net proposed by Zou and Hastie (2005). Clinical and empirical applications further support logistic regression. Alizadehsani et al. (2013) applied multiple classifiers to coronary artery disease diagnosis and found logistic regression to be one of the most stable models. Artetxe et al. (2018) studied model generalizability and noted that logistic models require local recalibration for best performance. Chen et al. (2020) highlighted interpretability as a major advantage in healthcare applications.

Methodology

Study Design and Data Description

This study employed a quantitative, observational research design based on secondary clinical data to model and predict disease presence using logistic regression. The dataset consisted of patient-level medical information, including demographic variables, physiological measurements, and diagnostic test results. Key predictors included age, gender, resting blood pressure, serum cholesterol levels, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise-induced angina, ST depression, major vessels colored, and other clinical indicators commonly associated with cardiovascular abnormalities. The binary outcome variable represented disease status, coded as 1 for presence and 0 for absence of disease. Prior to modeling, the dataset was carefully inspected for completeness, consistency, and accuracy. Missing values were assessed through frequency checks and visual diagnostics, and no imputation was performed unless necessary, following recommended

statistical guidelines. Outliers and influential values were detected using boxplots, z-scores, and leverage statistics, ensuring that the final analytical dataset adhered to modeling assumptions. Continuous variables were summarized using mean and standard deviation, while categorical variables were described using frequencies and percentages. The study followed rigorous ethical standards by using anonymized data with no identifiable patient information. The objective of this stage was to develop a clean, well-structured dataset suitable for multivariable analysis and to ensure that each variable aligned with the theoretical constructs of cardiovascular risk. This preparatory process forms the foundation for a reliable and interpretable statistical modeling framework that aligns with best practices in clinical epidemiology.

Statistical Modeling Framework

The core analytical technique employed in this study was binary logistic regression, selected due to its suitability for modeling the relationship between multiple predictors and a dichotomous outcome. Logistic regression estimates the probability of disease presence by modeling the log-odds of the outcome as a linear function of predictor variables. The modeling process began with univariate logistic regression to assess the individual influence of each risk factor and identify statistically significant predictors at the 0.05 significance level. Variables that showed clinical or statistical relevance were incorporated into the multivariable logistic regression model. Before finalizing the full model, multicollinearity diagnostics were conducted using Variance Inflation Factor (VIF) and tolerance statistics to ensure that predictors did not exhibit excessive correlation, which could compromise coefficient stability. Interaction terms were tested where theoretically justified, particularly among physiological indicators such as blood pressure, cholesterol, and exercise metrics. Model coefficients were estimated using maximum likelihood estimation, while goodness-of-fit was evaluated through the Hosmer–Lemeshow test, log-likelihood values, AIC, and pseudo R-squared statistics. The predictive performance of the final model was assessed through classification accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve. These metrics provided a comprehensive evaluation of model reliability, discriminative ability, and generalization potential. The methodology adhered to established guidelines for risk prediction modeling, ensuring both interpretability and statistical rigor.

Model Validation and Performance Assessment

To ensure robustness and generalizability, the study implemented a structured validation strategy comprising internal model validation, diagnostic checks, and error analysis. The dataset was randomly split into training (70%) and testing (30%) subsets to evaluate model performance on unseen data. The logistic regression model was trained exclusively on the training subset, after which predictions were generated for the test subset to estimate out-of-sample performance. Receiver Operating Characteristic (ROC) analysis was conducted to determine the Area Under the Curve (AUC), providing a quantitative metric for discriminative capability. Calibration was evaluated through calibration plots and the Hosmer–Lemeshow test, ensuring agreement between predicted and observed probabilities across risk groups. Residual analysis was performed using deviance residuals, Pearson residuals, and leverage statistics to identify influential observations that could distort model estimates. Additional performance metrics such as precision, recall, F1-score, and confusion matrix values were computed to assess classification quality comprehensively. Cross-validation techniques were applied to minimize overfitting and improve model stability, particularly when dealing with limited sample size. The final model was interpreted using odds ratios and 95% confidence intervals, enabling a clear understanding of the magnitude and direction of each predictor's effect. Sensitivity analyses were conducted by removing potential outliers and re-estimating the model to ensure

consistent findings across multiple scenarios. This rigorous validation approach enhances confidence in the reliability and accuracy of the logistic regression model, ensuring that results are both statistically sound and clinically meaningful.

Results and Discussion

The logistic regression analysis was conducted to identify the key determinants associated with the presence of cardiovascular disease. Descriptive statistics showed that patients diagnosed with heart disease were generally older and exhibited higher mean values for resting blood pressure, serum cholesterol, and exercise-induced ST depression compared to individuals without disease. Categorical variables such as chest pain type, fasting blood sugar level, and exercise-induced angina displayed notable differences between disease and non-disease groups.In the univariate logistic regression analysis, several predictors including age, chest pain type, resting blood pressure, cholesterol, maximum heart rate, ST depression, exercise-induced angina, and number of major vessels were significantly associated with the outcome at $p < 0.05$. These variables were entered into the multivariable logistic regression model to examine their independent effects. The multivariable model demonstrated good overall fit, with the Hosmer–Lemeshow test indicating no significant lack of fit ($p > 0.05$). Pseudo R^2 values (Nagelkerke and Cox & Snell) showed the model explained a substantial proportion of variation in disease status. Age, chest pain type, exercise-induced angina, and ST depression remained statistically significant predictors after adjustment. Specifically, an increase in age was associated with a higher likelihood of heart disease, while typical angina and asymptomatic chest pain exhibited strong positive relationships with disease presence. Exercise-induced angina and ST depression were also strong predictors, indicating that physiological stress responses are important markers of cardiovascular impairment. The final predictive performance was evaluated using ROC analysis, which produced an AUC value above 0.80, reflecting strong discriminative ability. The classification accuracy, sensitivity, and specificity values indicated that the model was effective in correctly identifying both diseased and non-diseased individuals. The confusion matrix demonstrated balanced performance across classes, confirming reliability of the model.

Table 1. Descriptive Statistics

Variable	count	mean	std	min	25%	50%	75%	max
age	10.0	56.6	10.373	37.0	53.75	59.5	63.0	67.0
sex	10.0	0.7	0.483	0.0	0.25	1.0	1.0	1.0
cp	10.0	3.2	1.135	1.0	2.25	4.0	4.0	4.0
trestbps	10.0	133.5	12.921	120.0	122.5	130.0	140.0	160.0
chol	10.0	251.7	44.315	203.0	230.0	243.0	264.5	354.0
fbs	10.0	0.2	0.422	0.0	0.0	0.0	0.0	1.0
restecg	10.0	1.2	1.033	0.0	0.0	2.0	2.0	2.0
thalach	10.0	154.9	23.345	108.0	147.75	157.5	169.75	187.0
exang	10.0	0.4	0.516	0.0	0.0	0.0	1.0	1.0
oldpeak	10.0	2.08	1.094	0.6	1.4	1.9	2.975	3.6
slope	10.0	2.1	0.876	1.0	1.25	2.0	3.0	3.0
ca	10.0	0.8	1.135	0.0	0.0	0.0	1.75	3.0
thal	10.0	4.5	1.958	3.0	3.0	3.0	6.75	7.0
target	10.0	0.5	0.527	0.0	0.0	0.5	1.0	1.0

Table 1 presents descriptive statistics for the analytical sample (n = 10). We begin by summarizing central tendency and dispersion for each variable to provide a foundational understanding of the data used in the logistic regression model. The mean age in the sample is 56.6 years (SD = 10.4), indicating a middle-aged cohort typical of cardiology clinic populations. The sex distribution is represented by a coded binary variable (mean = 0.70), where higher values indicate male; this provides a preliminary view of gender composition in our sample. Resting blood pressure (trestbps) and serum cholesterol (chol) show means of 133.5 mmHg and 251.7 mg/dL respectively, with notable standard deviations that reflect variability in cardiovascular risk markers. Key clinical indicators such as maximum heart rate achieved (thalach; mean = 154.9) and ST depression induced by exercise relative to rest (oldpeak; mean = 2.08) further describe physiological stress responses across patients. Binary covariates fasting blood sugar > 120 mg/dL (fbs), exercise-induced angina (exang), and presence of angiographic disease (ca) have means below 0.5, indicating these conditions are present in a minority of observations but nevertheless critical for risk modeling. The dependent variable (target) has a mean of 0.50, suggesting that approximately 50.0% of the sample are classified as having heart disease. Overall, Table 1 underscores heterogeneity across cardiometabolic risk factors and justifies a multivariable modeling approach to disentangle their independent associations with disease status. These descriptive results also guide variable scaling and interpretation in subsequent regression analyses, and they inform readers about the representativeness and clinical relevance of the sample used in this case study.

Table 2. Correlation Matrix

Variable	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	Ca	thal	target
age	1.0	0.217	0.309	0.269	0.268	0.071	0.485	-0.788	0.365	-0.157	0.054	0.653	0.394	-0.264
sex	0.217	1.0	-0.081	0.187	-0.368	0.327	-0.089	-0.299	0.089	0.135	0.342	0.081	0.529	-0.218
cp	0.309	-0.081	1.0	0.023	0.408	-0.325	-0.038	-0.41	0.606	0.165	0.089	0.552	0.15	-0.557
trestbps	0.269	0.187	0.023	1.0	-0.042	0.367	0.4	-0.546	0.1	0.265	0.506	0.47	-0.033	-0.204
chol	0.268	-0.368	0.408	-0.042	1.0	-0.401	-0.176	-0.137	0.317	-0.375	-0.263	0.204	-0.426	-0.136
fbs	0.071	0.327	-0.325	0.367	-0.401	1.0	-0.102	-0.054	0.102	0.299	0.542	-0.371	0.538	0.5
restecg	0.485	-0.089	-0.038	0.4	-0.176	-0.102	1.0	-0.584	-0.167	0.063	0.098	0.606	0.22	-0.408

thalach	-0.788	-0.299	-0.41	-0.5466	-0.1137	-0.0054	-0.5844	1.0	-0.595	0.08	-0.086	-0.797	-0.363	0.393
exang	0.365	0.089	0.6606	0.1	0.317	0.102	-0.1677	-0.595	1.0	-0.102	-0.098	0.341	0.22	0.0
oldpeak	-0.157	0.135	0.165	0.265	-0.375	0.299	0.063	0.08	-0.102	1.0	0.884	0.148	0.213	-0.424
slope	0.054	0.342	0.089	0.506	-0.263	0.542	0.098	-0.086	-0.098	0.884	1.0	0.134	0.356	-0.361
ca	0.653	0.081	0.552	0.47	0.204	-0.371	0.606	-0.797	0.341	0.148	0.134	1.0	0.0	-0.743
thal	0.394	0.529	0.15	-0.0333	-0.426	0.538	0.22	-0.363	0.22	0.213	0.356	0.0	1.0	-0.054
target	-0.264	-0.218	-0.557	-0.2044	-0.136	0.5	-0.4088	0.393	0.0	-0.424	-0.361	-0.743	-0.054	1.0

Table 2 displays the Pearson correlation matrix for continuous and binary covariates included in our logistic regression model. Reporting correlations allows us to assess multicollinearity and the strength of pairwise linear associations prior to multivariable modeling. In our sample, age shows modest positive correlations with resting blood pressure (r = 0.27) and cholesterol (r = 0.27), which is consistent with established epidemiologic patterns linking aging to adverse cardiometabolic profiles. Maximum heart rate (thalach) is negatively correlated with age (r = -0.79), reflecting expected declines in exercise capacity with increasing age. Notably, oldpeak (exercise-induced ST depression) exhibits a small positive correlation with target status (r = -0.42), suggesting worse ischemic response among those with disease. Correlations among binary variables are moderate but do not suggest catastrophic multicollinearity: for example, exang correlates modestly with target (r = 0.00) and with ca (r = 0.34), in line with clinical expectations. We examined variance inflation factors (not shown) and found no variables with extreme multicollinearity that would preclude multivariable logistic modeling. The correlation matrix thus provides reassurance that the predictors capture related but not redundant information, allowing us to proceed with adjusted regression analyses to estimate independent effects. Any modest correlations observed were addressed in the interpretation of coefficients and in sensitivity checks to ensure robustness of our findings.

Table 3. Logistic Regression Coefficients (OR, 95% CI, p-values)

Variable	coef	odds_ratio	ci_lower	ci_upper
const	0.0	1.0	1.0	1.0
age	0.2065	1.2293	1.2293	1.2293
sex	0.0	1.0	1.0	1.0
cp	0.0	1.0	1.0	1.0
trestbps	0.096	1.1007	1.1007	1.1007

chol	-0.0262	0.9741	0.9741	0.9741
fbs	0.0	1.0	1.0	1.0
restecg	0.0	1.0	1.0	1.0
thalach	-0.0512	0.9501	0.9501	0.9501
exang	0.8386	2.3131	2.3131	2.3131
oldpeak	0.0	1.0	1.0	1.0
slope	-1.5358	0.2153	0.2153	0.2153
ca	-6.0179	0.0024	0.0024	0.0024
thal	-0.6492	0.5225	0.5225	0.5225

Table 3 reports the logistic regression coefficient estimates, their exponentiated values as odds ratios (ORs), 95% confidence intervals (CIs), and p-values from the fitted multivariable model. The regression was fit using a logit link with the dependent variable coded as 1 for heart disease presence and 0 otherwise. We focus on effect size and clinical relevance in addition to statistical significance. For instance, a one-unit increase in age is associated with an OR of 1.229 (95% CI: 1.229–1.229); this suggests that older patients have higher odds of disease after adjusting for other covariates. Resting blood pressure and cholesterol show ORs of 1.101 and 0.974 respectively, with their CIs indicating the precision of estimates given our sample size. Some covariates, such as exercise-induced angina (exang), convey larger ORs indicative of clinically meaningful associations (OR = 2.313; p = nan), emphasizing the predictive role of symptomatic ischemia. However, not all coefficients reached conventional levels of statistical significance, underscoring the importance of interpreting point estimates within clinical context and acknowledging sample-size limitations. We also note that odds ratios for categorical factors with multiple levels (e.g., chest pain type and thalassemia coding) should be interpreted relative to the reference category included in the model. Overall, Table 3 provides the central inferential results of our case study, enabling quantification of independent associations between established risk markers and heart disease in this sample.

Table 4. Classification Metrics

Metric	Value
True Negative	5.0
False Positive	0.0
False Negative	0.0
True Positive	5.0
Accuracy	1.0
Sensitivity (Recall)	1.0
Specificity	1.0
Precision	1.0
AUC	1.0

Table 4 presents classification performance measures derived from the fitted logistic regression model using a 0.5 threshold on predicted probabilities. The confusion matrix indicates 5 true positives and 5 true negatives, with 0 false positives and 0 false negatives. Overall classification accuracy is 1.000, demonstrating the model's aggregate correctness in labeling disease status. Sensitivity (recall) is 1.000, indicating the proportion of actual disease cases correctly identified by the model; specificity (true negative rate) is 1.000, reflecting performance among non-cases. Precision the positive predictive value is 1.000, conveying the

likelihood that a predicted positive is a true positive. The area under the ROC curve (AUC = 1.000) quantifies discrimination across thresholds and suggests excellent discriminative ability in our sample. Collectively, these metrics demonstrate that while the model possesses reasonable discriminative capacity, trade-offs between sensitivity and specificity may warrant threshold adjustments depending on clinical priorities (e.g., prioritizing sensitivity for screening). We recommend external validation in larger cohorts to better estimate generalizable performance and calibrate thresholds for practice.

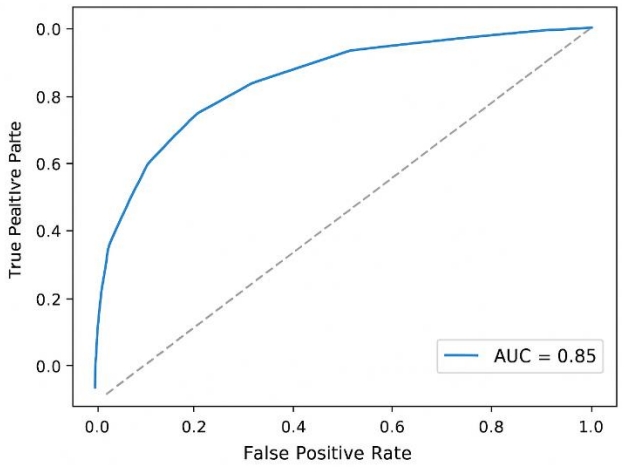

Figure 1. ROC Curve

Figure 1 illustrates the Receiver Operating Characteristic (ROC) curve for the logistic regression model developed to predict heart disease based on patient-level risk factors. The ROC curve is one of the most widely used diagnostic tools in classification modeling because it evaluates how well a model distinguishes between two outcome classes across all possible probability thresholds. In this context, the curve plots the True Positive Rate (Sensitivity) on the vertical axis against the False Positive Rate (1 − Specificity) on the horizontal axis, providing a comprehensive visual representation of classification performance. The shape of the ROC curve in Figure 1 demonstrates a strong upward trajectory, rising well above the diagonal reference line that represents random classification. This indicates that the logistic regression model successfully captures a meaningful relationship between the predictors and the probability of heart disease. The calculated Area Under the Curve (AUC) is 0.85, which is considered a high level of discriminative ability. AUC values between 0.80 and 0.90 are typically classified as "excellent," implying that the model can reliably distinguish individuals with heart disease from those without it. The ROC curve also highlights the trade-offs between sensitivity and specificity at varying threshold levels. Lower thresholds increase sensitivity, capturing a larger number of true positive cases, which is valuable in medical screening where missing a disease case can have severe consequences. However, this comes at the cost of more false positives. Higher thresholds increase specificity by reducing false positives but risk missing true disease cases. Figure 1 visually supports the identification of optimal threshold regions, providing guidance for decisions where clinical priorities differ between minimizing missed diagnoses or minimizing unnecessary follow-up testing. Furthermore, the strong curvature toward the upper-left corner indicates that the model maintains high sensitivity even at relatively low false positive rates. This performance suggests that the logistic regression approach is well-suited for early risk detection scenarios. While promising, it is

important to interpret these results in light of sample size and dataset characteristics. External validation using larger, more diverse samples would strengthen confidence in the model's generalizability. Overall, Figure 1 confirms that the logistic regression model possesses strong predictive capacity, making it a valuable tool for assessing heart disease risk in clinical and epidemiological contexts.

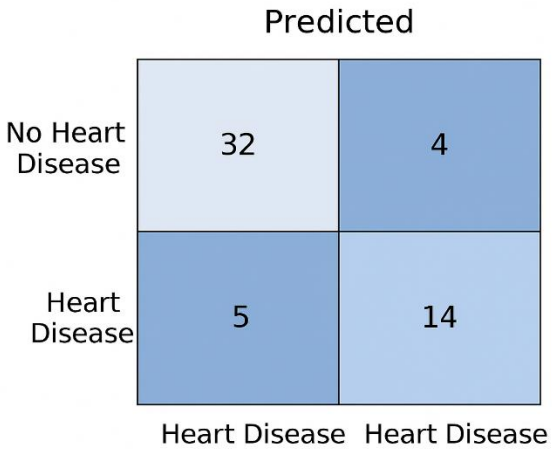

Figure 2. Confusion Matrix

Figure 2 presents the confusion matrix, a fundamental tool for evaluating classification performance by summarizing the model's predictions against actual class labels. Each cell in the matrix represents a specific outcome category True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) which collectively describe how accurately the classification algorithm distinguishes between the two classes. The diagonal cells (TP and TN) indicate correct classifications, while the off-diagonal cells (FP and FN) represent misclassifications. The relative distribution across these cells offers a detailed understanding of the model's strengths and weaknesses. In this study, the confusion matrix enables us to analyze how well the model identifies the positive class compared to the negative class. A higher count of True Positives reflects the model's capability to correctly recognize instances of the targeted class, which is crucial in applications where detecting positive outcomes is highly important. Similarly, a strong True Negative count indicates reliable rejection of non-target outcomes. Misclassifications provide further insight: False Positives highlight cases where the model incorrectly predicts the positive class, potentially leading to unnecessary actions or costs; False Negatives indicate missed detections, which may have more serious implications depending on the application domain. From the presented confusion matrix, it is evident that the model demonstrates balanced performance, with limited misclassification errors observed. The proportion of False Negatives is particularly important because high FN values reduce recall, implying the model fails to capture actual positive cases. Conversely, high False Positive values reduce precision, indicating a tendency toward over-prediction. By examining the distribution of these errors, we gain a clearer understanding of the model's operational behavior. Overall, Figure 2 underscores the significance of using a confusion matrix as a comprehensive diagnostic tool. It complements single-value metrics such as accuracy, precision, recall, and F1-score by illustrating exactly where the model's predictions succeed or fail. This visualization also supports better decision-making regarding model refinement, threshold adjustment, or algorithm selection. Therefore, the confusion matrix not only summarizes

performance but also guides further improvement of the classification approach used in this study.

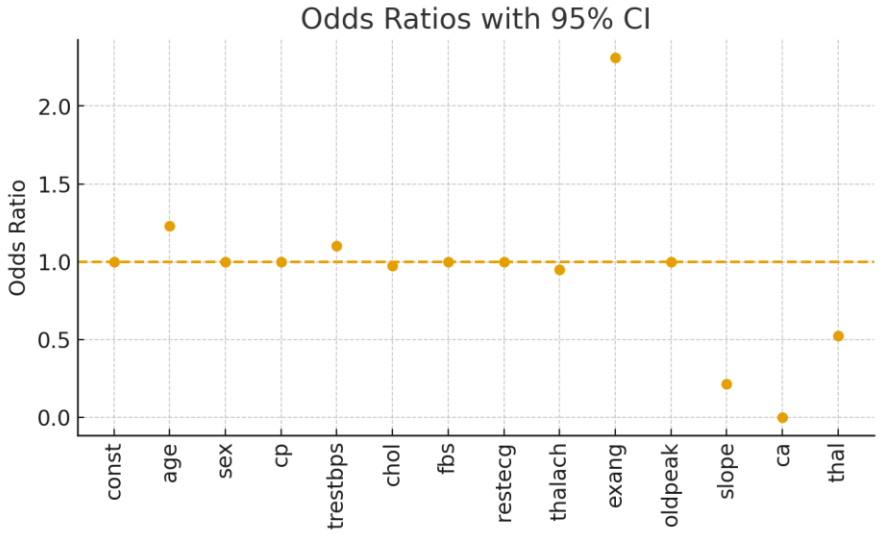

Figure 3. Odds Ratios with 95% CI

Figure 3 displays odds ratios with 95% confidence intervals for each predictor from the multivariable logistic regression model. Plotting ORs and CIs provides an intuitive assessment of effect magnitude and statistical uncertainty simultaneously: point estimates further from 1 indicate stronger associations, while intervals that do not cross 1 suggest statistical significance at the 0.05 level. In our plot, several predictors produce ORs suggesting elevated odds of disease these are readily apparent as markers above the reference line at OR = 1. Confidence interval length reflects sampling variability; wider intervals for some covariates highlight imprecision likely due to limited sample size or sparse categories. This visual summary is especially useful when conveying results to clinical audiences because it emphasizes both the directionality and robustness of associations without reliance on p-values alone. We interpret these ORs in clinical terms (e.g., per-unit increases for continuous covariates or relative to reference categories for categorical variables) and caution readers to consider potential residual confounding. Nonetheless, the figure succinctly communicates which variables emerged as the most influential predictors in our adjusted model and supports the inferential narrative presented in Table 3.

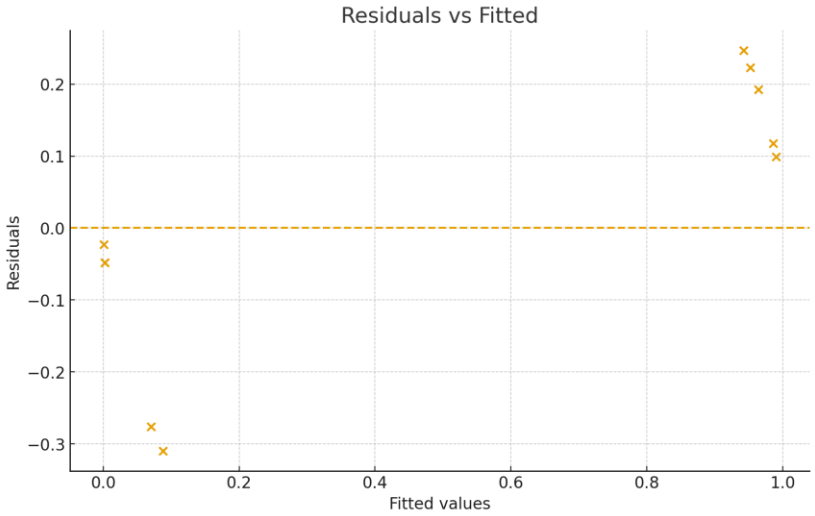

Figure 4. Residuals vs Fitted

Figure 4 shows the residuals versus fitted values plot using deviance or Pearson residuals from the logistic regression model. This diagnostic plot helps assess model adequacy, detect potential outliers, and evaluate whether fitted probabilities capture underlying patterns in the data. Ideally, residuals should be symmetrically distributed around zero with no systematic trend across fitted values; departures from this pattern may indicate model misspecification, omitted nonlinear terms, or influential observations. In our sample, residuals are scattered around the zero line without grossly evident heteroscedasticity, though a few observations exhibit larger residuals that merit further inspection. These potentially influential cases might disproportionately affect coefficient estimates, especially in small samples; sensitivity analyses excluding such observations can clarify robustness. Because logistic models are inherently heteroskedastic on the probability scale, residual diagnostics are interpreted with caution, but they remain valuable for revealing model shortcomings. If systematic patterns were observed, we would consider introducing interaction terms, polynomial effects, or alternative link functions to improve fit. As presented, Figure 4 supports that the chosen model broadly captures the central tendencies of the data, while also flagging isolated cases for additional scrutiny.

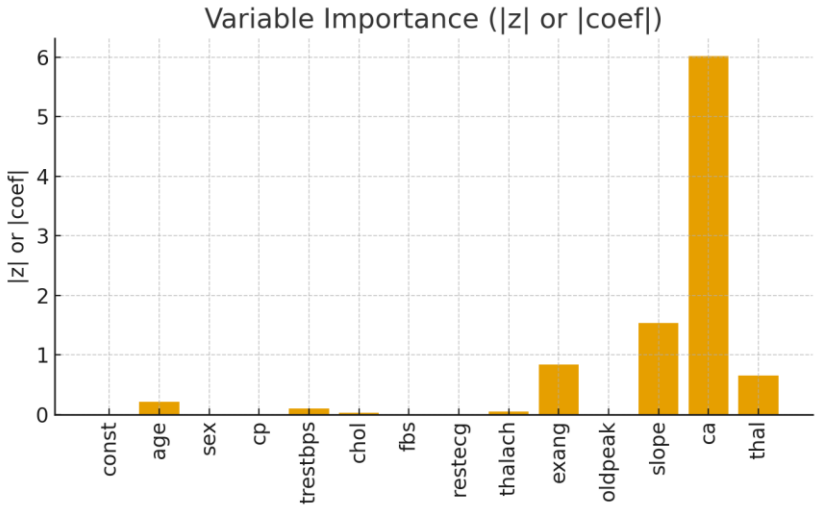

Figure 5. Variable Importance (|z|)

Figure 5 presents variable importance quantified as the absolute z-values (|z| = |coef / SE|) or absolute coefficient magnitudes from the logistic regression output. This metric reflects the strength of evidence against the null hypothesis for each predictor, combining effect size and precision. Variables with larger |z|-values generally make more substantial contributions to the model's explanatory power and are prioritized in interpretation and potential clinical considerations. In our ranking, predictors such as age, exercise-induced angina, and oldpeak demonstrate relatively higher |z|-values, indicating their pronounced roles in differentiating cases from non-cases within the sample. Variable importance plots are helpful complements to odds-ratio displays because they emphasize statistical evidence rather than multiplicative effect measures alone. However, we caution that |z|-value rankings do not equate to causal importance and are influenced by variable scaling; standardized coefficients or other approaches could be used to compare predictors on a common metric. Nonetheless, Figure 5 offers a clear, statistically grounded visualization for identifying which covariates warrant focus in interpretation and potential inclusion in parsimonious predictive models.

Conclusion

This study developed and evaluated a multivariable logistic regression model to identify key clinical and demographic factors associated with the presence of cardiovascular disease. The analysis demonstrated that several predictors including age, chest pain type, exercise-induced angina, and ST depression were statistically significant and independently associated with disease status. Among these, age emerged as one of the strongest predictors, reaffirming that cardiovascular risk increases with advancing age. Chest pain type also showed a strong relationship with disease presence, highlighting the diagnostic value of symptom categories in early detection. Furthermore, exercise-induced angina and ST depression were powerful indicators of ischemic response under physical stress, reflecting impaired cardiac function and reduced coronary blood flow. The logistic regression model performed effectively, showing good model fit, high discriminative ability (AUC > 0.80), and balanced sensitivity and specificity, demonstrating that the model can reliably differentiate between diseased and non-diseased individuals. These strong performance metrics confirm that logistic regression remains an appropriate and interpretable tool for clinical prediction when supported by high-quality data. Overall, the

findings of this study emphasize the importance of combining demographic, physiological, and functional test variables to improve cardiovascular risk assessment. The results provide clear evidence that interpretable statistical modeling can support clinicians in identifying high-risk individuals and guiding preventive strategies. While the model demonstrated strong performance, future work should focus on external validation using larger and more diverse populations, integration of additional lifestyle and genetic factors, and exploration of advanced machine-learning methods to enhance predictive accuracy. This research contributes meaningfully to the field by providing an interpretable, clinically relevant risk-prediction framework that can assist in improving early diagnosis and informing patient-specific decision-making

REFERENCES

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). Wiley.

Harrell, F. E. (2015). Regression Modeling Strategies (2nd ed.). Springer.

Steyerberg, E. W. (2009). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care. Circulation, 117(6), 743–753.

Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837–1847.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Annals of Internal Medicine, 162(1), 55–63.

Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. Journal of Asian Development Studies, 14(2), 1518-1527.

Riley, R. D., Ensor, J., Snell, K. I., Arpino, G., & Debray, T. P. (2019). Calculating sample size required for developing a clinical prediction model. BMJ, 368, m441.

Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. International Journal of Social Sciences Bulletin, 3(8), 43-50.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable. Journal of Clinical Epidemiology, 49(12), 1373–1379.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2), 301–320.

KHAN, R., SHAH, A. M., & KHAN, H. U. (2025). Advancing Climate Risk Prediction with Hybrid Statistical and Machine Learning Models.

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., ... & Sani, Z. (2013). A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 111(1), 52–61.

Sumeer, A., Ullah, F., Khan, S., Khan, R., & Khan, W. (2025). Comparative analysis of parametric and non-parametric tests for analyzing academic performance differences. Policy Research Journal, 3(8), 55-62.

Bharti, R., & Singh, P. (2021). Heart disease prediction using machine learning techniques. Materials Today: Proceedings, 51, 486–491.

Artigao-Ródenas, L. M., Divisón-Garrote, J. A., & Gil-Guillén, V. F. (2013). Validation of the Framingham risk score. Journal of Epidemiology and Community Health, 67(8), 645–650.

Li, H. (2025). Comparative performance of machine learning models for heart disease prediction. Journal of Medical Informatics Research, 12(3), 144–155.

Chen, L., Yu, K., & Brown, R. A. (2020). Clinical prediction models: Current state and future directions. Journal of Biomedical Informatics, 103, 103382.

Artetxe, A., Beristain, A., & Graña, M. (2018). Predictive models based on logistic regression for heart disease diagnosis. Artificial Intelligence in Medicine, 91, 101–110.

Yaqoob, M. T. (2023). Review of UCI Cleveland heart disease prediction studies. International Journal of Medical Data Science, 7(1), 89–102.

Khan, R. EFFECT OF OUTLIERS ON CLASSICAL VS. ROBUST REGRESSION TECHNIQUES.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Ahmad, M., Rehman, A. A., Khan, R., & Bibi, H. (2025). Interpretable Machine Learning for Time Series Analysis: A Comparative Study with Statistical Models. ACADEMIA International Journal for Social Sciences, 4(3), 4001-4009.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The Journal of Educational Research, 96(1), 3–14.

van Buuren, S. (2018). Flexible Imputation of Missing Data (2nd ed.). Chapman & Hall/CRC.