ISSN 3007-3170(O)
ISSN :3007-3162(P)

January-March 2025

Social Sciences & Humanity Research Review

Heart Disease Risk PredictionStatistical Analysis and Classification of Heart Diseases Risk Using Clinical Parameter

Bushra Shehzadi[1], Hafiz Abdul Sami[2], Dr. Shahbaz Nawaz[3], Dr. Anam Javaid[4], Tooba Nihal[5], Fatima Bibi[6]

[1]Department of Statistics ,The Women University Multan, Pakistan
[2]Bureau Of Statistics , Planning & Development Board, Govt of the Punjab, Pakistan
[3]Bureau Of Statistics, Planning & Development Board, Govt of the Punjab, Pakistan
[4] Assistant Professor , Department of Statistics , The Women University Multan, Pakistan
[5]Lecturer , Department of Statistics , The Women University Multan, Pakistan.
[6]Department of Statistics ,The Women University Multan, Pakistan

ARTICLE INFO

Keywords: Heart Disease Risk, Statistical Modeling, Logistic Regression, Machine Learning, Random Forest, Neural Networks

Corresponding Author:
Bushra Shehzadi,
Department of Statistics , The Women University Multan, Pakistan

ABSTRACT

Heart disease remains a leading cause of morbidity and mortality worldwide, necessitating robust statistical methods for early detection and risk prediction. This study applies multiple statistical and machine learning techniques, including Logistic Regression, Random Forest, and Neural Networks, to analyze a clinical dataset of 1,025 observations with 14 variables related to demographic, physiological, and biochemical parameters. The study evaluates model performance using accuracy, sensitivity, specificity, and AUC metrics, while also assessing multicollinearity, correlation structures, and variable significance through standardized coefficients and information criteria (AIC/BIC). The Logistic Regression model achieved an AUC of 0.83, indicating strong predictive capability, whereas ensemble methods required parameter tuning to improve specificity. The results highlight key risk factors including chest pain type, ST depression, maximum heart rate, and number of major vessels, offering data-driven insights for preventive healthcare strategies.

1. Introduction

Cardiovascular diseases (CVDs), particularly heart disease, are among the leading causes of morbidity and mortality worldwide. According to the World Health Organization (WHO), an estimated 17.9 million people die each year from CVDs, accounting for nearly one-third of global deaths. Early detection and accurate classification of individuals at risk are critical to reducing the burden of these diseases and improving population health outcomes.

Statistical modeling and machine learning techniques provide powerful tools to analyze complex relationships between clinical parameters and disease outcomes. By examining demographic, physiological, and biochemical variables simultaneously, researchers can identify key risk factors and develop predictive models for early diagnosis.

In recent years, techniques such as logistic regression, Random Forest, and Neural Networks have been increasingly applied in medical research to improve diagnostic accuracy and support clinical decision-making. However, there is a need for comparative studies evaluating the performance and interpretability of these methods in heart disease risk prediction.

This study aims to bridge this gap by performing a comprehensive statistical analysis and classification of heart disease risk using multiple clinical parameters. The findings offer insights into the most significant predictors and provide recommendations for integrating statistical and machine learning approaches into preventive healthcare strategies.

2. Literature Review

Previous studies have explored risk factors such as hypertension, cholesterol levels, diabetes, and behavioral aspects using regression and survival models (Poulter, 1999; Haffner, 2000). Machine learning approaches, including Random Forests and Neural Networks, have shown promise in capturing non-linear relationships between variables (Carney & Freedland, 2017). However, comparisons between classical statistical models and modern machine learning algorithms remain limited in heart disease prediction contexts.

3. Methodology

The dataset, sourced from the Kaggle Heart Disease repository, comprises 1,025 observations across 14 variables including age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, slope of ST segment, number of major vessels, and thalassemia type. Descriptive statistics summarize variable distributions, followed by correlation analysis and multicollinearity assessment using Variance Inflation Factor (VIF). Predictive models include Logistic Regression, Random Forest, and Neural Networks. Model selection relies on accuracy, AUC, sensitivity, specificity, and information criteria (AIC/BIC).The best-performing model was selected based on higher AUC and balanced sensitivity-specificity performance.

4. Results and Discussion

Descriptive analysis revealed a mean age of 54.4 years with 51.3% of participants diagnosed with heart disease. Correlation analysis identified chest pain type, maximum heart rate, and slope as positively associated with heart disease, while exercise-induced angina and ST depression exhibited negative correlations. Logistic Regression yielded significant predictors ($p<0.05$) including chest pain, ST depression, number of vessels, and maximum heart rate, achieving an AUC of 0.83. Random Forest models demonstrated high sensitivity but low specificity, suggesting overfitting risks without parameter tuning. Neural Networks underperformed due to limited training data and parameter optimization needs.

Descriptive statistics of Various Factor

Variable	Observations	Obs. with missing data	Obs. without missin	Minimum	Maximum	Mean	Std. deviation

g data							
target	1025	0	1025	0.000	1.000	0.513	0.500
age	1025	0	1025	29.000	77.000	54.434	9.072
sex	1025	0	1025	0.000	1.000	0.696	0.460
cp	1025	0	1025	0.000	3.000	0.942	1.030
trestbps	1025	0	1025	94.000	200.000	131.612	17.517
chol	1025	0	1025	126.000	564.000	246.000	51.593
fbs	1025	0	1025	0.000	1.000	0.149	0.357
restecg	1025	0	1025	0.000	2.000	0.530	0.528
thalach	1025	0	1025	71.000	202.000	149.114	23.006
exang	1025	0	1025	0.000	1.000	0.337	0.473
oldpeak	1025	0	1025	0.000	6.200	1.072	1.175
slope	1025	0	1025	0.000	2.000	1.385	0.618
ca	1025	0	1025	0.000	4.000	0.754	1.031

Correlation matrix

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	**1**	-0.103	-0.072	0.271	0.220	0.121	-0.133	-0.390	0.088	0.208	-0.169	0.272	0.072	-0.229
sex	-0.103	**1**	-0.041	-0.079	-0.198	0.027	-0.055	-0.049	0.139	0.085	-0.027	0.112	0.198	-0.280
cp	-0.072	-0.041	**1**	0.038	-0.082	0.079	0.044	0.307	-0.402	-0.175	0.132	-0.176	-0.163	0.435

trestbps	0.271	-0.079	0.038	**1**	0.128	0.182	-0.124	-0.039	0.061	0.187	-0.120	0.105	0.059	-0.139
chol	0.220	-0.198	-0.082	0.128	**1**	0.027	-0.147	-0.022	0.067	0.065	-0.014	0.074	0.100	-0.100
fbs	0.121	0.027	0.079	0.182	0.027	**1**	-0.104	-0.009	0.049	0.011	-0.062	0.137	-0.042	-0.041
restecg	-0.133	-0.055	0.044	-0.124	-0.147	-0.104	**1**	0.048	-0.066	-0.050	0.086	-0.078	-0.021	0.134
thalach	-0.390	-0.049	0.307	-0.039	-0.022	-0.009	0.048	**1**	-0.380	-0.350	0.395	-0.208	-0.098	0.423
exang	0.088	0.139	-0.402	0.061	0.067	0.049	-0.066	-0.380	**1**	0.311	-0.267	0.108	0.197	-0.438
oldpeak	0.208	0.085	-0.175	0.187	0.065	0.011	-0.050	-0.350	0.311	**1**	-0.575	0.222	0.203	-0.438
slope	-0.169	-0.027	0.132	-0.120	-0.014	-0.062	0.086	0.395	-0.267	-0.575	**1**	-0.073	-0.094	0.346
ca	0.272	0.112	-0.176	0.105	0.074	0.137	-0.078	-0.208	0.108	0.222	-0.073	**1**	0.149	-0.382
thal	0.072	0.198	-0.163	0.059	0.100	-0.042	-0.021	-0.098	0.197	0.203	-0.094	0.149	**1**	-0.338

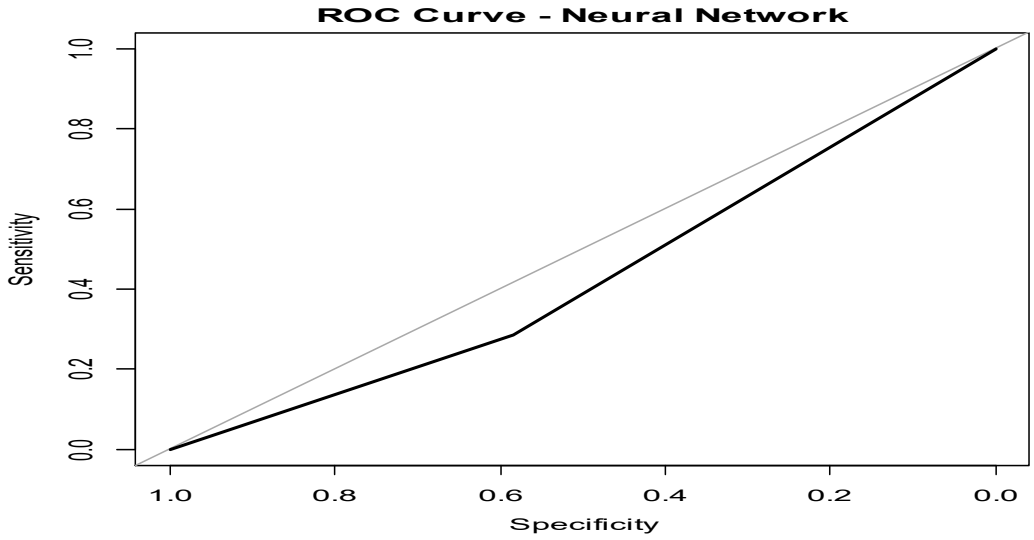
target	-0.229	-0.280	0.435	-0.139	-0.100	-0.041	0.134	0.423	-0.438	-0.438	0.346	-0.382	-0.338	**1**

Multicollinearity statistics

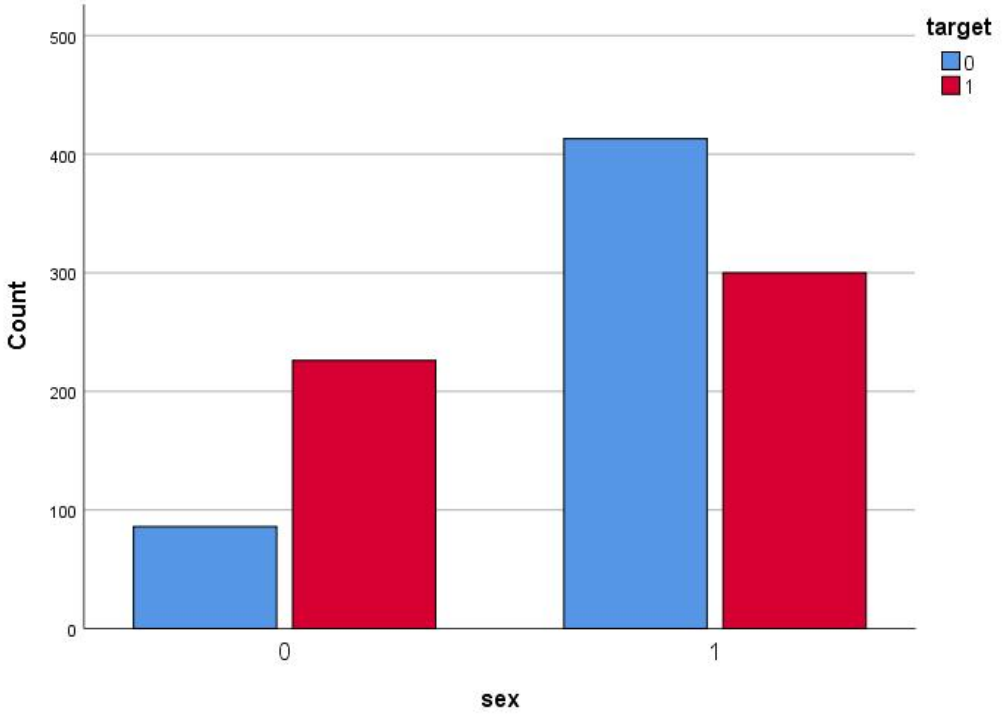
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
Tolerance	**0.700**	**0.865**	**0.773**	**0.856**	**0.873**	**0.917**	**0.939**	**0.619**	**0.705**	**0.585**	**0.609**	**0.835**	**0.879**
VIF	**1.429**	**1.156**	**1.293**	**1.168**	**1.146**	**1.090**	**1.064**	**1.615**	**1.419**	**1.709**	**1.643**	**1.197**	**1.138**

Standardised coefficient

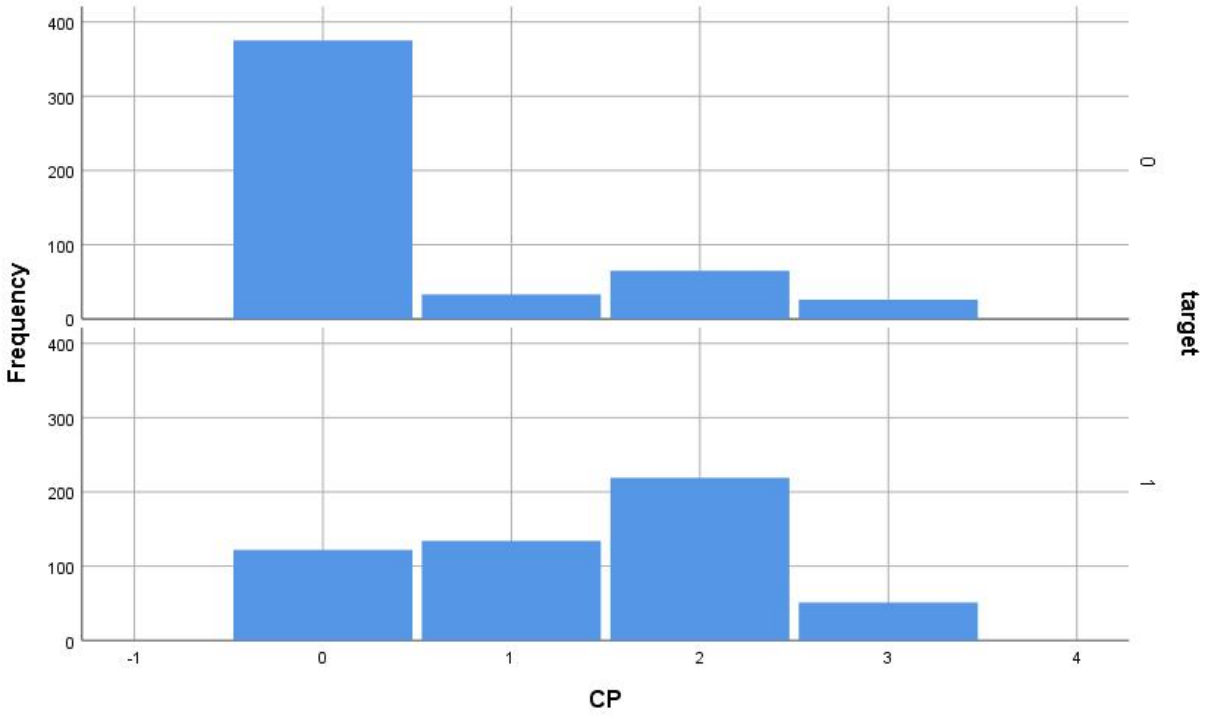
Source	Value	Standard error	Wald Chi-Square	Pr > Chi²	Wald Lower bound (95%)	Wald Upper bound (95%)
age	-0.041	0.063	0.423	0.516	-0.165	0.083
sex	-0.468	0.065	51.797	<0.0001	-0.596	-0.341
cp	0.485	0.057	72.521	<0.0001	0.373	0.596
trestbps	-0.176	0.054	10.529	0.001	-0.282	-0.070
chol	-0.161	0.058	7.603	0.006	-0.276	-0.047
fbs	-0.020	0.056	0.126	0.723	-0.130	0.090
restecg	0.120	0.055	4.781	0.029	0.012	0.228
thalach	0.300	0.072	17.292	<0.0001	0.158	0.441
exang	-0.258	0.058	19.514	<0.0001	-0.373	-0.144
oldpeak	-0.370	0.075	24.206	<0.0001	-0.517	-0.222
slope	0.182	0.064	8.012	0.005	0.056	0.308
ca	-0.429	0.059	53.603	<0.0001	-0.543	-0.314
thal	-0.303	0.053	32.415	<0.0001	-0.407	-0.199



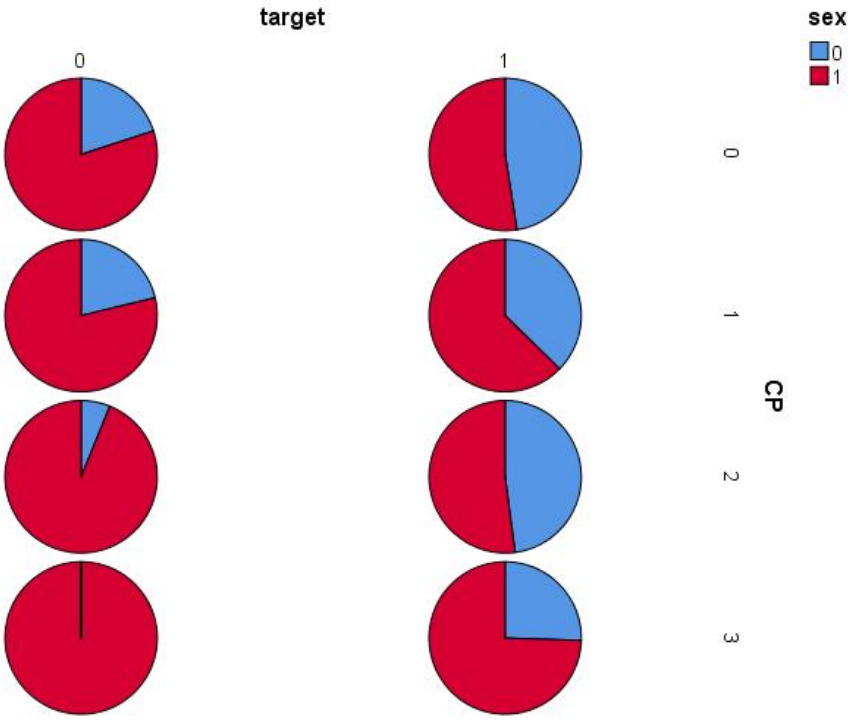
Bar Chart



Histogram



Pie Chart



5. Conclusion and Recommendations

Statistical and machine learning techniques offer complementary strengths in heart disease risk prediction. Logistic Regression provides interpretable results with strong discriminatory power, while Random Forest and Neural Networks require further optimization for clinical deployment. Future research should focus on larger datasets, feature engineering, and ensemble learning to enhance specificity and overall predictive accuracy.

References

Poulter, N. R. (1999). Coronary heart disease is a multifactorial disease. *European Heart Journal*, 20(9), 685-692.
Haffner, S. M. (2000). Coronary heart disease in patients with diabetes. *New England Journal of Medicine*, 342(14), 1040-1042.
Carney, R. M., & Freedland, K. E. (2017). Depression and coronary heart disease. *Nature Reviews Cardiology*, 14(3), 145-155.
Barker, d. J. 1995. Fetal origins of coronary heart disease. Bmj, 311, 171-174.
Carney, r. M. & freedland, k. E. 2017. Depression and coronary heart disease. Nature reviews cardiology, 14, 145-155.

CHANDOLA, T., BRITTON, A., BRUNNER, E., HEMINGWAY, H., MALIK, M., KUMARI, M., BADRICK, E., KIVIMAKI, M. & MARMOT, M. 2008. Work stress and coronary heart disease: what are the mechanisms? European heart journal, 29, 640-648.
CHEN, Y., WANG, L., MA, D., CUI, Z., LIU, Y., PANG, Q., JIANG, Z. & GAO, Z. 2024. Research on rheumatic heart disease from 2013 to early 2024: a bibliometric analysis. Journal of Cardiothoracic Surgery, 19, 659.
DANESH, J., COLLINS, R. & PETO, R. 1997. Chronic infections and coronary heart disease: is there a link? The lancet, 350, 430-436.
DAWBER, T. R., MOORE, F. E. & MANN, G. V. 2015. II. Coronary heart disease in the Framingham study. International journal of epidemiology, 44, 1767-1780.
DRAZNER, M. H. 2011. The progression of hypertensive heart disease. Circulation, 123, 327-334.
GRUNDY, S. M. 1986. Cholesterol and coronary heart disease: a new era. Jama, 256, 2849-2858.
HAFFNER, S. M. 2000. Coronary heart disease in patients with diabetes. Mass Medical Soc.
HAINES, A., IMESON, J. & MEADE, T. 1987. Phobic anxiety and ischaemic heart disease. Br Med J (Clin Res Ed), 295, 297-299.

HANCOCK, S. L., TUCKER, M. A. & HOPPE, R. T. 1993. Factors affecting late mortality from heart disease after treatment of Hodgkin's disease. Jama, 270, 1949-1955.
HOFFMAN, J. I. & KAPLAN, S. 2002. The incidence of congenital heart disease. Journal of the American college of cardiology, 39, 1890-1900.
HOFFMAN, J. I., KAPLAN, S. & LIBERTHSON, R. R. 2004. Prevalence of congenital heart disease. American heart journal, 147, 425-439.
KEYS, A. 1957. Diet and the epidemiology of coronary heart disease. Journal of the American Medical Association, 164, 1912-1919.
KOENIG, W. 2001. Inflammation and coronary heart disease: an overview. Cardiology in review, 9, 31-35.
LIU, J. L., MANIADAKIS, N., GRAY, A. & RAYNER, M. 2002. The economic burden of coronary heart disease in the UK. Heart, 88, 597-603.
LLOYD-JONES, D. M., LARSON, M. G., BEISER, A. & LEVY, D. 1999. Lifetime risk of developing coronary heart disease. The Lancet, 353, 89-92.
MCCARTHY, M., LAY, M. & ADDINGTON-HALL, J. 1996. Dying from heart disease. Journal of the Royal College of Physicians of London, 30, 325.
NASCHITZ, J. E., SLOBODIN, G., LEWIS, R. J., ZUCKERMAN, E. & YESHURUN, D. 2000. Heart diseases affecting the liver and liver diseases affecting the heart. American heart journal, 140, 111-120.

PAUL, O., LEPPER, M. H., PHELAN, W. H., DUPERTUIS, G. W., MACMILLAN, A., MCKEAN, H. & PARK, H. 1963. A longitudinal study of coronary heart disease. Circulation, 28, 20-31.
PEARSON, T. A. 1996. Alcohol and heart disease. Circulation, 94, 3023-3025.
PELL, J. & COBBE, S. 1999. Seasonal variations in coronary heart disease. Qjm, 92, 689-696.
POULTER, N. 1999. Coronary heart disease is a multifactorial disease. American Journal of Hypertension, 12, 92S-95S.
UEBING, A., STEER, P. J., YENTIS, S. M. & GATZOULIS, M. A. 2006. Pregnancy and congenital heart disease. Bmj, 332, 401-406.
VIRANI, S. S., ALONSO, A., APARICIO, H. J., BENJAMIN, E. J., BITTENCOURT, M. S., CALLAWAY, C. W., CARSON, A. P., CHAMBERLAIN, A. M., CHENG, S. & DELLING, F. N. 2021. Heart disease and stroke statistics—2021 update. Circulation.