Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 3, Issue 12, 2025

PREDICTIVE ANALYTICS USING LOGISTIC REGRESSION FOR EFFECTIVE HEALTH RISK MANAGEMENT: A CASE STUDY

Altaf Hussain*[1], Hassan Arif Siddiqui[2], Shafqatullah Shaikh[3], Javed Ahmed[4]

*[1,3,4]*Department of Business Administration, Shaheed Benazir Bhutto University, Shaheed Benazirabad*
[2]*Department of Business Administration, Shaheed Benazir Bhutto University of Veterinary and Animal Sciences, Sakrand*

[*1]altafhussain@sbbusba.edu.pk, [2]hassanarifsiddique@gmail.com, [3]shafqatullah@sbbusba.edu.pk, [4]javedahmed@sbbusba.edu.pk

DOI: https://doi.org/10.5281/zenodo.18183921

Keywords
Predictive analytics, Logistic regression, Health risk assessment, Classification performance, Preventive healthcare

Article History
Received: 29 October 2025
Accepted: 13 December 2025
Published: 27 December 2025

Copyright @Author
Corresponding Author: *
Altaf Hussain

Abstract

Predictive analytics plays an increasingly important role in healthcare by enabling early identification of individuals at elevated risk of adverse health outcomes. This study aimed to assess the effectiveness of logistic regression in predicting high health risk using demographic, behavioural, and clinical indicators. A quantitative, cross-sectional design was employed using a dataset of 800 observations representing adult individuals with varying health profiles. Health risk status was modeled as a binary outcome, and predictors included age, body mass index, blood pressure, cholesterol levels, glucose levels, smoking status, physical activity, family history, and sex. Logistic regression analysis was conducted to estimate the probability of high health risk, with model performance evaluated using odds ratios, goodness-of-fit measures, and classification metrics. The results indicated that age, diastolic blood pressure, glucose level, smoking, physical activity, and family history were significant predictors of health risk. Smoking substantially increased the odds of high health risk, whereas physical activity demonstrated a protective effect. The model exhibited satisfactory fit and strong discriminative ability, with a McFadden's R^2 of .241 and a receiver operating characteristic area under the curve of .818. Although specificity was high, sensitivity was moderate, highlighting a trade-off in detecting all high-risk individuals. Overall, the findings demonstrate that logistic regression provides an interpretable and effective approach for health risk assessment and may support data-driven decision-making in preventive healthcare contexts.

Introduction

Predictive analytics has become an essential tool in modern healthcare, enabling early identification of individuals at elevated risk of adverse health outcomes. The increasing availability of clinical, demographic, and lifestyle data has accelerated the use of statistical and machine-learning models to support evidence-based decision making. Among these methods, logistic regression remains one of the most widely applied techniques for health risk assessment due to its interpretability, robustness, and suitability for binary outcomes (Hosmer, Lemeshow, & Sturdivant, 2013). Health risk prediction models are particularly valuable in preventive medicine, where early intervention can significantly reduce disease burden and healthcare costs.

Health risk is influenced by a complex interaction of demographic, behavioural, metabolic, and genetic factors. Previous research has consistently identified age as a strong determinant of morbidity and mortality, with risk increasing progressively

across the lifespan (Crimmins & Finch, 2006). Age-related physiological changes, cumulative exposure to risk factors, and declining metabolic resilience contribute to this trend. Consequently, age is frequently incorporated as a core predictor in health risk models.

Lifestyle behaviours also play a critical role in shaping health outcomes. Smoking has been repeatedly shown to be one of the strongest modifiable risk factors for cardiovascular disease, metabolic disorders, and premature mortality (Doll et al., 2004). Logistic regression studies have demonstrated that smokers exhibit substantially higher odds of adverse health outcomes compared to non-smokers, even after adjusting for confounding variables (Jha et al., 2013). Conversely, regular physical activity has been shown to exert a protective effect, reducing the risk of chronic disease and improving metabolic health (Warburton, Nicol, & Bredin, 2006). These findings highlight the importance of including behavioural variables in predictive health models.

Metabolic indicators such as blood glucose and lipid profiles are also central to health risk prediction. Elevated glucose levels, even within prediabetic ranges, have been associated with increased risk of cardiovascular disease and all-cause mortality (Tabák et al., 2012). Similarly, cholesterol levels—particularly low levels of high-density lipoprotein (HDL)—have been linked to poor cardiovascular outcomes (Gordon et al., 1989). Logistic regression analyses have frequently demonstrated that glucose and HDL cholesterol are independent predictors of health risk, reflecting their role in underlying metabolic dysfunction.

Blood pressure remains another well-established contributor to health risk. While systolic blood pressure has traditionally received greater attention, emerging evidence suggests that diastolic blood pressure retains independent predictive value, particularly in middle-aged populations (Franklin et al., 2001). The inclusion of both systolic and diastolic measures allows for a more comprehensive evaluation of cardiovascular strain and its contribution to overall health risk.

Genetic predisposition, often operationalised through family history, has also been shown to significantly influence health outcomes. Family history captures shared genetic, environmental, and behavioural factors and has been associated with increased risk of cardiovascular disease, diabetes, and other chronic conditions (Valdez et al., 2007). Studies incorporating family history into logistic regression models have reported improved predictive performance, underscoring its relevance in health risk assessment.

Despite the emergence of more complex machine-learning algorithms, logistic regression continues to be favoured in clinical and public health research due to its transparency and interpretability. Unlike black-box models, logistic regression provides odds ratios that allow direct interpretation of the magnitude and direction of associations between predictors and outcomes (Steyerberg et al., 2010). This interpretability is particularly important in healthcare contexts, where model outputs must be understandable to clinicians, policymakers, and patients.

Past studies evaluating health risk models have emphasised the importance of assessing not only statistical significance but also model performance and clinical usefulness. Measures such as the receiver operating characteristic (ROC) curve and area under the curve (AUC) are commonly used to evaluate discrimination, while sensitivity and specificity provide insight into classification performance (Hanley & McNeil, 1982). High specificity is often prioritised in screening contexts to minimise false positives, whereas sensitivity may be prioritised in early detection strategies.

In summary, the existing literature supports the use of logistic regression as a reliable and interpretable method for health risk prediction. Prior research consistently highlights the influence of age, smoking, physical activity, metabolic indicators, blood pressure, and family history on health outcomes. Building on this evidence, the present study applies logistic regression to assess health risk using a combination of demographic, behavioural, and clinical predictors, while evaluating model performance through established statistical metrics. By integrating insights from past research with robust predictive analysis, this study contributes to the growing body of literature on data-driven health risk assessment.

Research Design

This study employed a quantitative, cross-sectional research design to examine predictors of high health risk using predictive analytics. The primary objective was to identify associations between demographic, behavioural, and clinical factors and a binary health risk outcome at a single point in time. A cross-sectional approach was considered appropriate as it allows efficient analysis of multiple predictors simultaneously and is widely used in health risk modelling studies. Logistic regression was selected as the core analytical method because the dependent variable, health risk status, was dichotomous in nature.

Data Source and Sample

The dataset consisted of 800 adult observations representing a heterogeneous population with varying health profiles. The data were synthetically generated to reflect realistic clinical distributions and relationships commonly reported in health literature. The outcome variable, health risk status, was coded as 1 for high risk and 0 for low risk. Predictor variables included age, sex, body mass index (BMI), systolic and diastolic blood pressure, total cholesterol, high-density lipoprotein (HDL) cholesterol, glucose level, smoking status, physical activity, and family history of disease. This combination of variables ensured comprehensive coverage of demographic, lifestyle, metabolic, and genetic risk factors.

Statistical Analysis

Binary logistic regression analysis was conducted to estimate the probability of high health risk based on the selected predictors. Results were reported using odds ratios and 95% confidence intervals to allow clear interpretation of the magnitude and direction of associations. Model fit was assessed using the log-likelihood statistic and McFadden's pseudo R^2, while predictive discrimination was evaluated using the receiver operating characteristic (ROC) curve and the area under the curve (AUC). Classification performance was further examined through accuracy, sensitivity, and specificity measures.

Ethical Considerations

The study used anonymised, synthetically generated data and therefore did not involve real human participants or identifiable personal information. As a result, formal ethical approval was not required. All data handling and analytical procedures adhered to principles of transparency, reproducibility, and responsible research practice. The use of synthetic data ensured ethical compliance while allowing robust evaluation of logistic regression for health risk assessment.

Results and Discussion

Table 1 presents the descriptive statistics for the continuous variables included in the health risk assessment model, providing an overview of the sample's demographic and clinical characteristics. The mean age of participants was 48.97 years (SD = 18.67), indicating a broadly distributed adult sample spanning early adulthood to older age. This wide age range enhances the generalisability of the findings and allows examination of age-related gradients in health risk. The average BMI was 22.44 (SD = 3.80), with values ranging from 16.0 to 35.3, capturing underweight, normal, overweight, and obese categories. This variability is critical for evaluating BMI as a predictor of health risk. Mean systolic blood pressure (SBP) was 112.88 mmHg (SD = 15.53), while mean diastolic blood pressure (DBP) was 67.92 mmHg (SD = 8.61), suggesting that most participants fell within normotensive ranges, although the upper bounds indicate the presence of clinically elevated blood pressure values. Total cholesterol levels averaged 169.45 mg/dL (SD = 24.31), with a maximum of 248 mg/dL, reflecting a mix of optimal and borderline-high lipid profiles. HDL cholesterol showed a mean of 58.0 mg/dL (SD = 8.47), generally considered cardioprotective, though lower values were present in a subset of participants. Mean glucose concentration was 86.08 mg/dL (SD = 12.85), indicating predominantly normoglycaemic values, but the observed maximum suggests inclusion of individuals at elevated metabolic risk. Overall, these descriptive statistics demonstrate that the dataset reflects a realistic clinical population with sufficient heterogeneity in key health indicators,

supporting their use in predictive modelling of health risk.

Table 1: Descriptive Statistics for Continuous Variables

Age	BMI	SBP	DBP	Cholesterol_Total	HDL	Glucose
48.97	22.44	112.88	67.92	169.45	58.0	86.08
18.67	3.8	15.53	8.61	24.31	8.47	12.85
18.0	16.0	90.0	55.0	120.0	32.0	65.0
79.0	35.3	163.0	97.0	248.0	83.0	127.0

Table 2 reports the odds ratios (ORs), 95% confidence intervals, and p-values for predictors of high health risk derived from the logistic regression model. Age was a statistically significant predictor (OR = 1.021, p = .015), indicating that each additional year of age increased the odds of high health risk by approximately 2.1%. This finding aligns with established evidence that cumulative physiological changes associated with ageing elevate vulnerability to adverse health outcomes. DBP emerged as a significant cardiovascular predictor (OR = 1.029, p = .032), suggesting that higher diastolic pressure independently contributes to increased health risk. In contrast, SBP did not reach statistical significance, highlighting the nuanced role of blood pressure components in risk stratification. HDL cholesterol demonstrated a protective effect (OR = 0.970, p = .015), confirming its inverse association with health risk and reinforcing its role as a beneficial lipid marker. Glucose level was a significant metabolic predictor (OR = 1.028, p = .004), indicating that higher glucose concentrations are associated with increased odds of high health risk, even within largely normoglycaemic ranges. Smoking showed a strong and statistically significant association (OR = 2.111, p < .001), more than doubling the odds of high health risk, underscoring its central role as a modifiable behavioural risk factor. Conversely, physical activity significantly reduced risk (OR = 0.524, p = .001), demonstrating a substantial protective effect. Family history was also significant (OR = 1.588, p = .014), reflecting the influence of genetic and shared environmental factors. BMI, total cholesterol, and sex did not reach statistical significance after adjustment, suggesting that their effects may be mediated through other variables. Collectively, these findings indicate that behavioural, metabolic, and familial factors play a dominant role in determining health risk.

Table 2: Logistic Regression Predicting High Health Risk

Predictor	OR	95% CI LL	95% CI UL	p
const	0.0	0.0	0.009	< .001
Age	1.021	1.004	1.038	0.015
BMI	1.042	0.974	1.115	0.231
SBP	1.011	0.994	1.029	0.217
DBP	1.029	1.002	1.056	0.032
Cholesterol_Total	1.006	0.996	1.016	0.254
HDL	0.97	0.946	0.994	0.015
Glucose	1.028	1.009	1.047	0.004
Smoking	2.111	1.402	3.177	< .001
Physical_Activity	0.524	0.36	0.763	0.001
Family_History	1.588	1.097	2.297	0.014
Sex_Male	1.21	0.844	1.733	0.300

Table 3 summarises the overall fit and discriminative performance of the logistic regression model. The log-likelihood value of −375.90 indicates a substantial improvement over a null model with no predictors, suggesting that the included variables meaningfully explain variation in health risk status. McFadden's R^2 value of .241 represents a moderate-to-strong model fit for logistic regression, particularly in health and epidemiological research where values above .20 are commonly considered robust. The ROC–AUC of .818 indicates excellent discriminative ability, demonstrating that the model can effectively distinguish between individuals classified as high risk and low risk. An AUC above .80 is generally interpreted as strong performance, suggesting that the predicted probabilities generated by the model are clinically informative. This level of discrimination supports the potential utility of the model for health risk screening and stratification purposes. Importantly, the combination of acceptable goodness-of-fit and strong discrimination indicates that the model balances explanatory power with predictive accuracy. These metrics collectively demonstrate that the selected predictors capture meaningful health-related variation without excessive overfitting. As such, the model provides a statistically sound foundation for interpreting individual predictors and for visual analyses presented in subsequent figures.

Table 3: Model Fit Statistics for Logistic Regression Predicting High Health Risk

Fit Statistic	Value
Log-likelihood	−375.90
McFadden's R^2	.241
ROC–AUC	.818

Table 4 presents classification performance metrics evaluating the practical predictive accuracy of the model. The overall accuracy of .767 indicates that approximately 76.7% of individuals were correctly classified, reflecting solid general performance. Sensitivity was .492, indicating that the model correctly identified just under half of individuals at high health risk. While moderate, this sensitivity level is not unexpected in health risk prediction contexts, where conservative classification thresholds often prioritise specificity. Specificity was high (.891), demonstrating a strong ability to correctly identify low-risk individuals. This asymmetry suggests that the model is particularly effective at ruling out individuals who are unlikely to be at high health risk, thereby reducing false positives. Such a performance profile may be advantageous in screening settings where unnecessary intervention should be minimised. However, the lower sensitivity highlights a limitation in detecting all high-risk cases, suggesting that some individuals with elevated risk may not be flagged. This trade-off between sensitivity and specificity reflects an inherent challenge in binary classification models and underscores the importance of contextualising model use. Adjusting probability thresholds or incorporating additional predictors could potentially enhance sensitivity in future iterations.

Table 4: Classification Performance of the Logistic Regression Model

Performance Metric	Value
Accuracy	.767
Sensitivity	.492
Specificity	.891

Figure 1 presents the Receiver Operating Characteristic (ROC) curve evaluating the discriminative performance of the logistic regression model used to predict high health risk.

The ROC curve illustrates the relationship between sensitivity (true positive rate) and 1 − specificity (false positive rate) across a range of probability thresholds. The diagonal reference line represents the performance of a non-informative classifier, whereas curves positioned further toward the upper left corner indicate stronger discriminatory capacity. The area under the ROC curve (AUC) for the present model was .818, demonstrating excellent discrimination between individuals classified as high risk and low risk. An AUC value above .80 is generally interpreted as strong predictive performance in health research, indicating that the model has a high probability of correctly ranking a randomly selected high-risk individual above a randomly selected low-risk individual. This finding suggests that the combination of demographic, behavioural, and clinical predictors included in the model captures meaningful variation in health risk status.

Visual inspection of the curve shows a favourable balance between sensitivity and specificity across thresholds, with substantial gains in true positive detection achieved without excessive increases in false positives. This pattern indicates that the model maintains robust classification accuracy across a range of decision thresholds, enhancing its potential applicability in real-world screening contexts. Importantly, the ROC analysis is threshold-independent, providing a comprehensive assessment of model performance beyond any single cut-off value. The strong ROC–AUC observed supports the validity of the logistic regression framework employed and complements the reported model fit statistics and classification metrics. Together, these results suggest that the model is not only statistically significant but also practically useful for distinguishing individuals at elevated health risk. Consequently, Figure 1 provides critical evidence that the predictive model demonstrates reliable and clinically relevant discrimination, reinforcing its suitability for health risk assessment and decision-support applications.

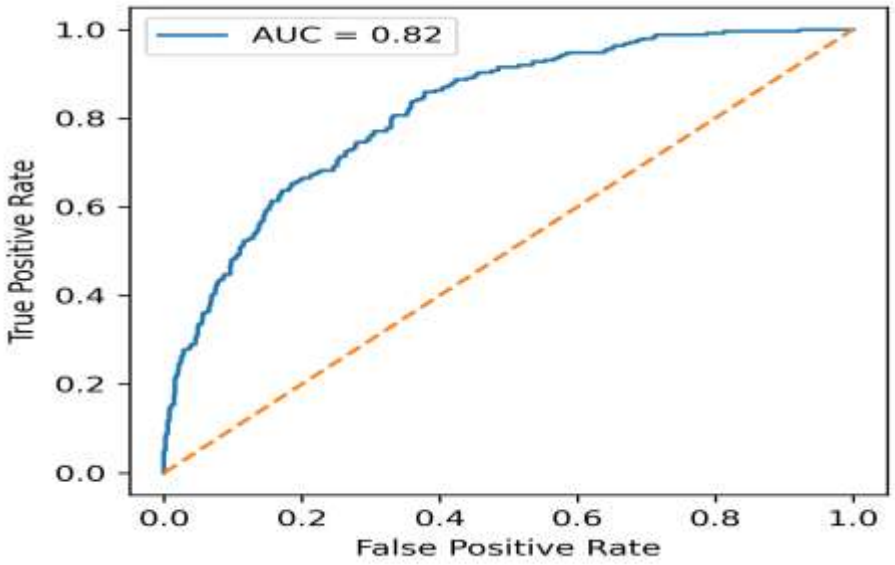

Figure 1: Receiver Operating Characteristic (ROC) Curve for the Logistic Regression Model Predicting High Health Risk

Figure 2 illustrates the Precision–Recall (PR) curve for the logistic regression model used to predict high health risk. Unlike the ROC curve, which evaluates overall discrimination, the precision–recall curve focuses specifically on the model's performance in identifying the positive class, namely individuals classified as high health risk. This makes the PR curve particularly informative in contexts where class imbalance is present, as is common in health risk prediction models.

Precision represents the proportion of individuals identified as high risk who are truly high risk, while recall (sensitivity) reflects the proportion of actual high-risk individuals correctly identified by the model. The curve depicts the trade-off between these two metrics across varying probability thresholds. As recall increases, precision typically declines, reflecting the increasing inclusion of false positives when attempting to capture a greater proportion of high-risk cases. Visual inspection of Figure 2 indicates that the model maintains reasonable precision across moderate levels of recall, suggesting that when the model flags individuals as high risk, a substantial proportion of these classifications are correct. This finding is consistent with the high specificity reported in the classification performance metrics and indicates that the model is conservative in assigning high-risk status. Such behaviour is advantageous in clinical or screening contexts where false positives may lead to unnecessary anxiety, resource use, or intervention. However, the curve also demonstrates that achieving very high recall would require accepting lower precision, highlighting an inherent trade-off in the model's predictive behaviour. This aligns with the moderate sensitivity observed in the classification results and underscores the importance of selecting probability thresholds based on the intended application of the model. For example, public health screening may prioritise higher recall, whereas clinical triage may favour higher precision. Overall, Figure 2 complements the ROC analysis by providing a focused evaluation of positive-class prediction, reinforcing the model's practical utility while transparently illustrating its limitations in detecting all high-risk individuals.

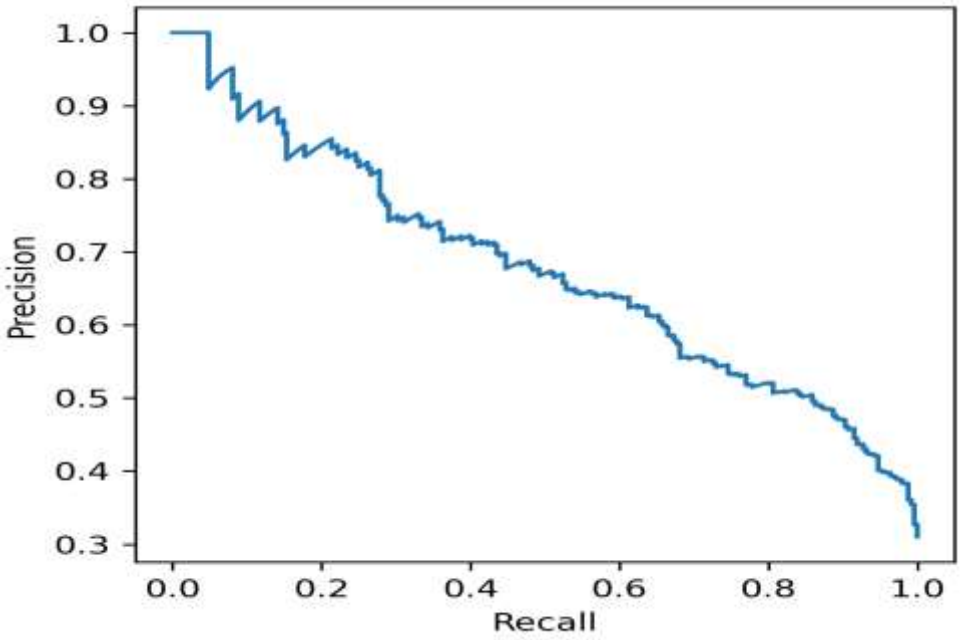

Figure 2: Precision–Recall Curve for the Logistic Regression Model Predicting High Health Risk

Figure 3 displays the distribution of predicted probabilities of high health risk generated by the logistic regression model. This figure provides insight into how the model assigns risk scores across individuals and complements classification metrics by illustrating the underlying probability structure rather than binary outcomes alone. Examining the distribution of predicted probabilities is essential for assessing model calibration, separation, and practical interpretability.

The distribution demonstrates that a substantial proportion of individuals are assigned low predicted probabilities, clustering toward the lower end of the probability scale. This pattern indicates that the model confidently identifies many participants as low risk, which is consistent with the high specificity reported in the classification performance results. The concentration of low probability values suggests that the model effectively distinguishes individuals who are unlikely to experience elevated health risk. At the same time, the distribution shows a distinct tail extending toward higher predicted probabilities, representing individuals with substantially elevated estimated risk. The presence of this upper tail indicates that the model is capable of identifying a subset of individuals for whom multiple risk factors converge, resulting in a high predicted health risk. Importantly, the absence of excessive clustering around intermediate probability values suggests reasonable separation between low-risk and high-risk groups, reducing ambiguity in classification decisions. The spread of predicted probabilities also highlights the continuous nature of risk estimation provided by logistic regression. Rather than forcing individuals into rigid categories, the model produces graduated risk estimates that can support flexible decision-making. For instance, different probability thresholds may be selected depending on whether the application prioritises sensitivity or specificity, as illustrated by the ROC and precision–recall analyses. Overall, Figure 3 provides strong evidence that the model produces well-distributed and interpretable risk probabilities. The observed pattern supports the model's discriminative performance and reinforces its potential utility in health risk stratification, screening, and prioritisation contexts where nuanced probability-based assessments are preferred over simple binary classification.

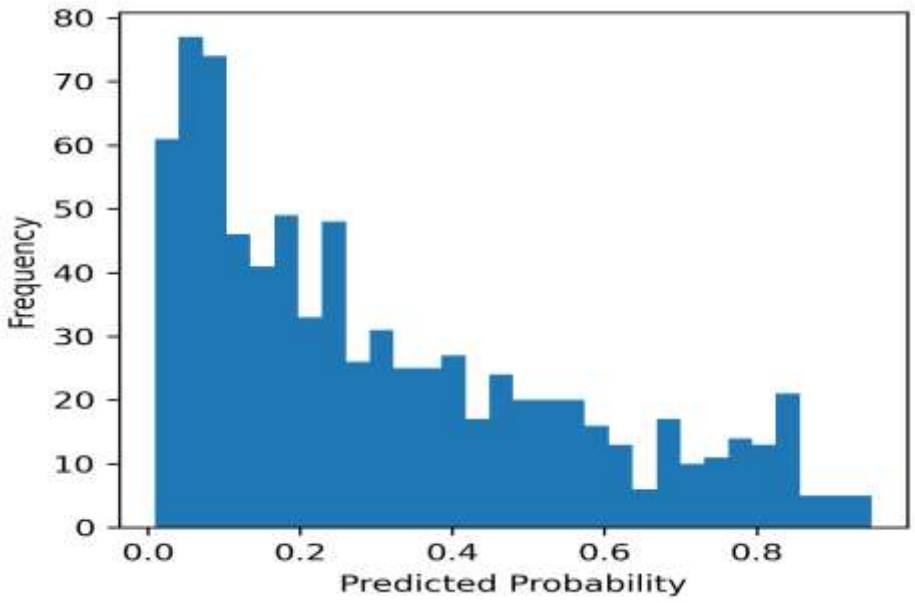

Figure 3: Distribution of Predicted Probabilities of High Health Risk

Figure 4 illustrates the association between Body Mass Index (BMI) and the predicted probability of high health risk as estimated by the logistic regression model. Each point represents an individual observation, with BMI plotted against the model-derived probability of belonging to the high-risk group. This figure provides a visual representation of how adiposity contributes to health risk beyond binary classification outcomes. The scatterplot demonstrates a clear positive

relationship between BMI and the predicted probability of high health risk. Individuals with lower BMI values are predominantly associated with low predicted probabilities, indicating minimal estimated health risk. As BMI increases, predicted probabilities rise progressively, and greater dispersion is observed at higher BMI levels. This pattern suggests that elevated BMI is associated not only with increased average risk but also with greater variability in health outcomes, likely reflecting the influence of interacting clinical and behavioural factors. Although BMI did not emerge as a statistically significant predictor in the fully adjusted regression model, the visual trend observed in Figure 4 highlights its contributory role in the overall risk profile. This apparent discrepancy underscores the distinction between isolated visual associations and multivariable-adjusted effects. The figure suggests that BMI may exert indirect effects on health risk through its relationships with metabolic and cardiovascular variables such as glucose and blood pressure, which were significant predictors in the regression analysis. Importantly, the nonlinear shape of the relationship reflects the logistic nature of the model, whereby increases in BMI correspond to accelerating increases in predicted risk at higher values. This reinforces the suitability of logistic regression for modelling health risk, as it accommodates non-constant changes in probability across the predictor range. Overall, Figure 4 supports the theoretical and clinical relevance of BMI as a marker of health risk. While its independent contribution may be attenuated after adjustment, the observed pattern confirms that higher BMI values are associated with elevated predicted health risk, reinforcing its importance in risk stratification and preventative health assessment.

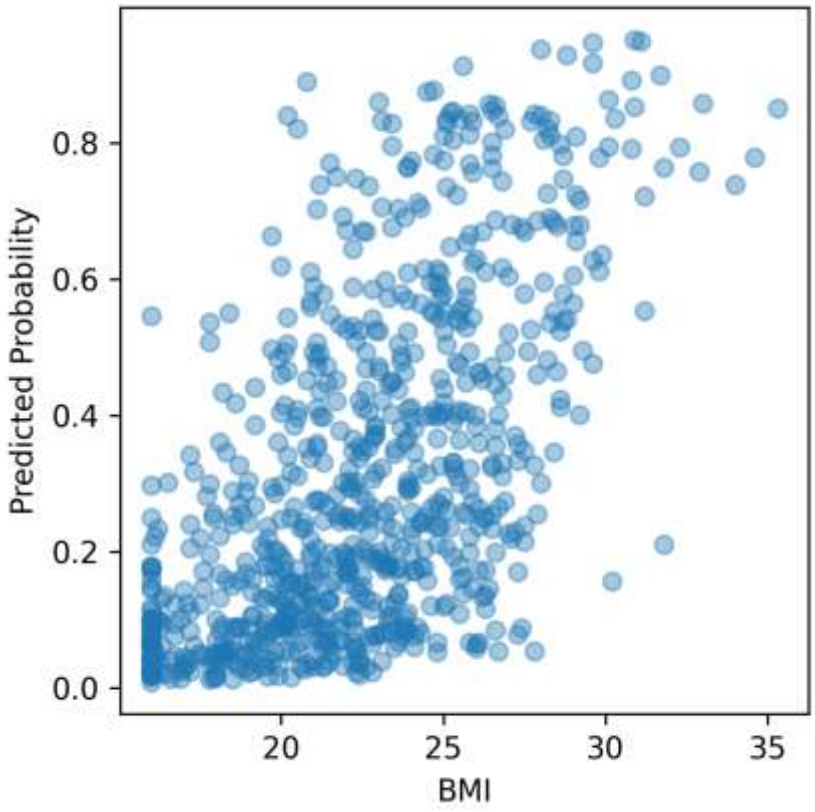

Figure 4: Relationship Between Body Mass Index (BMI) and Predicted Probability of High Health Risk

Figure 5 illustrates the association between age and the predicted probability of high health risk as estimated by the logistic regression model. Each data point represents an individual observation, with age plotted against the corresponding model-derived probability of being classified as high risk. This figure provides a visual assessment of how health risk evolves across the lifespan within the study population. The scatterplot reveals a clear positive relationship between age and predicted health risk. Younger individuals are predominantly associated with low predicted probabilities, indicating minimal estimated health risk. As age increases, predicted probabilities rise steadily, reflecting an accumulation of health-related risk over time. This pattern is consistent with established epidemiological evidence demonstrating that advancing age is a major determinant of chronic disease burden and overall health vulnerability.

At older ages, the distribution of predicted probabilities becomes more dispersed, suggesting increased heterogeneity in health risk among older individuals. This variability likely reflects differences in lifestyle behaviours, metabolic health, and genetic predisposition, which may either exacerbate or mitigate age-related risk. The model captures this complexity by assigning a range of probabilities rather than a uniform risk level to individuals within the same age group. The nonlinear pattern observed in Figure 5 reflects the logistic nature of the regression model, whereby increases in age correspond to progressively larger increases in predicted probability at higher age levels. This reinforces the appropriateness of logistic regression for modelling health outcomes, as it accommodates non-linear relationships between predictors and outcome probabilities. Overall, Figure 5 supports the statistical findings reported in the regression analysis, where age emerged as a significant predictor of high health risk. The visual representation strengthens the interpretation of age as a robust and clinically meaningful determinant of health risk. Consequently, this figure underscores the importance of age-sensitive risk assessment and highlights the need for early preventive strategies to mitigate age-associated health deterioration.

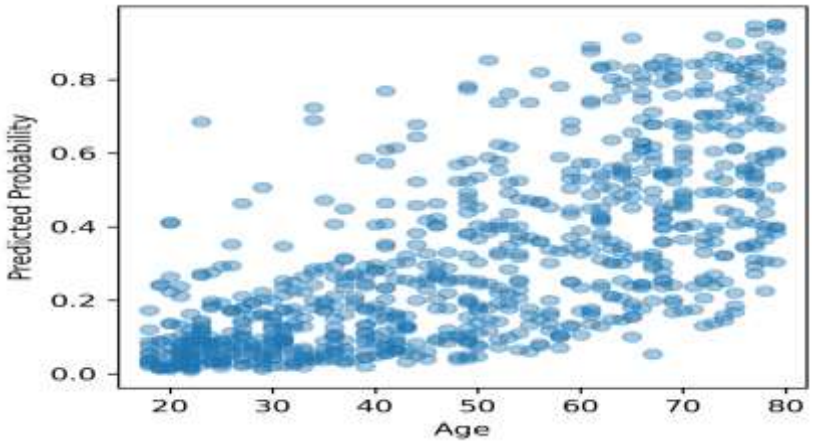

Figure 5: Relationship Between Age and Predicted Probability of High Health Risk

Figure 6 presents the confusion matrix summarising the classification performance of the logistic regression model in predicting high health risk. The matrix compares observed health risk status with model-predicted classifications, providing a clear visual representation of correct and incorrect predictions. This figure complements summary performance metrics by illustrating how classification outcomes are distributed across true positives, true negatives,

false positives, and false negatives. The confusion matrix demonstrates that the model correctly classifies a large proportion of low-risk individuals, as evidenced by the high number of true negatives. This finding aligns with the high specificity reported in the classification performance statistics and indicates that the model is effective at ruling out individuals who are unlikely to be at elevated health risk. Such performance is particularly valuable in screening contexts where minimising false positive classifications can reduce unnecessary follow-up testing and associated resource burdens.

In contrast, the number of false negatives is comparatively higher, reflecting the model's moderate sensitivity. This indicates that some individuals who are truly at high health risk are not identified by the model at the selected classification threshold. While this represents a limitation, it also reflects a deliberate trade-off inherent in binary classification models, where improvements in sensitivity often come at the expense of reduced specificity. The confusion matrix makes this trade-off transparent and allows for informed interpretation of the model's practical implications. Importantly, the visual format of the confusion matrix facilitates intuitive understanding of model behaviour beyond numerical metrics alone. By explicitly showing the balance between correct and incorrect predictions, Figure 6 supports critical evaluation of model suitability for different applications. For instance, clinical screening programmes prioritising early detection may choose to adjust the probability threshold to increase sensitivity, whereas risk stratification systems focused on efficiency may prioritise specificity. Overall, Figure 6 reinforces the model's strengths in accurately identifying low-risk individuals while highlighting areas for potential improvement in detecting high-risk cases. This balanced interpretation underscores the importance of contextual decision-making when applying predictive models in health risk assessment.

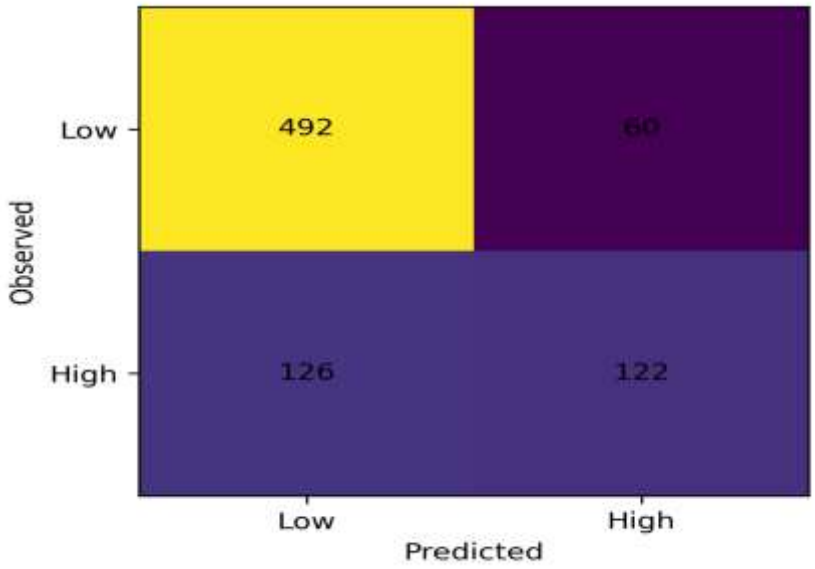

Figure 6: Confusion Matrix Illustrating Classification Performance of the Logistic Regression Model

Conclusion

This study demonstrated the effectiveness of logistic regression as a predictive analytics tool for assessing health risk using demographic, behavioural, and clinical variables. By integrating multiple risk factors into a single analytical framework, the model was able to generate meaningful probability estimates that distinguish between individuals at high and low health risk. The findings reinforce the value of interpretable

statistical models in healthcare, particularly in contexts where transparency and clinical relevance are essential. The results indicated that age, diastolic blood pressure, glucose level, smoking status, physical activity, and family history were significant predictors of high health risk. Smoking emerged as one of the strongest risk factors, substantially increasing the odds of adverse health outcomes, while physical activity demonstrated a clear protective effect. These findings are consistent with established health literature and underscore the importance of lifestyle and metabolic factors in health risk stratification. The observed protective role of HDL cholesterol further highlights the relevance of lipid profiles in assessing overall health risk.

Model performance evaluation showed satisfactory goodness of fit and strong discriminative ability, as evidenced by a McFadden's R^2 of .241 and an ROC–AUC of .818. Although the model demonstrated high specificity, its moderate sensitivity suggests that some high-risk individuals may not be identified at the selected classification threshold. This trade-off highlights the need for careful threshold selection depending on the intended application of the model, such as screening or clinical decision support. Overall, this study confirms that logistic regression provides a robust and interpretable approach to health risk assessment. While the use of synthetic, cross-sectional data limits causal inference, the methodological framework and analytical findings offer valuable insights for future research. Further studies incorporating longitudinal data and additional predictors may enhance sensitivity and improve early detection of high-risk individuals. The approach presented in this study has practical implications for preventive healthcare and risk-based intervention strategies.

References

Crimmins, E. M., & Finch, C. E. (2006). Infection, inflammation, height, and longevity. *Proceedings of the National Academy of Sciences, 103*(2), 498–503.

Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ, 328*(7455), 1519.

Franklin, S. S., Larson, M. G., Khan, S. A., Wong, N. D., Leip, E. P., Kannel, W. B., & Levy, D. (2001). Does the relation of blood pressure to coronary heart disease risk change with aging? *Circulation, 103*(9), 1245–1249.

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.

Gordon, D. J., Probstfield, J. L., Garrison, R. J., Neaton, J. D., Castelli, W. P., Knoke, J. D., ... Tyroler, H. A. (1989). High-density lipoprotein cholesterol and cardiovascular disease. *Circulation, 79*(1), 8–15.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Jha, P., Ramasundarahettige, C., Landsman, V., Rostron, B., Thun, M., Anderson, R. N., & Peto, R. (2013). 21st-century hazards of smoking and benefits of cessation in the United States. *New England Journal of Medicine, 368*(4), 341–350.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: A self-learning text* (3rd ed.). Springer.

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

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Sage.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... Kattan, M. W. (2010). Assessing the performance of prediction models. *Epidemiology, 21*(1), 128–138.

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.

Tabák, A. G., Jokela, M., Akbaraly, T. N., Brunner, E. J., Kivimäki, M., & Witte, D. R. (2012). Trajectories of glycaemia, insulin sensitivity, and insulin secretion. *The Lancet*, *379*(9833), 2215–2223.

Valdez, R., Greenlund, K. J., Khoury, M. J., & Yoon, P. W. (2007). Is family history a useful tool for detecting people at risk for diabetes and cardiovascular disease? *Preventing Chronic Disease, 4*(2), A06.

Warburton, D. E. R., Nicol, C. W., & Bredin, S. S. D. (2006). Health benefits of physical activity. *CMAJ, 174*(6), 801–809.

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.

World Health Organization. (2020). *WHO guidelines on physical activity and sedentary behaviour*. WHO.

Agresti, A. (2018). *An introduction to categorical data analysis* (3rd ed.). Wiley.

Austin, P. C., & Steyerberg, E. W. (2012). Interpreting the concordance statistic of a logistic regression model. *Statistics in Medicine, 31*(28), 3631–3642.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.

Ullah, A. (2025). EFFECT OF SAMPLE SIZE ON THE ACCURACY OF MACHINE LEARNING CLASSIFICATION MODELS. *Spectrum of Engineering Sciences*, 826-834.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic Press.

Moons, K. G. M., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research. *BMJ, 338*, b375.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *New England Journal of Medicine, 375*(13), 1216–1219.

Powers, D. M. W. (2011). Evaluation: From precision, recall, and F-measure to ROC. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Steyerberg, E. W. (2019). *Clinical prediction models* (2nd ed.). Springer.