Scientific Research Publishing

Voice of the Publisher, 2026, 12(2), 205-216
https://www.scirp.org/journal/vp
ISSN Online: 2380-7598
ISSN Print: 2380-7571

Revisiting Logistic Regression for Diabetes Prediction: An Interpretable and High-Performance Analysis Using a Public Health Dataset

Peter Chimwanda[1*], Edwin Rupi[2]

[1]Vice Chancellor's Office, Chinhoyi University of Technology, Chinhoyi, Zimbabwe
[2]Department of Mathematics, Masvingo Teachers College Masvingo, Masvingo, Zimbabwe
Email: *pchimwanda@cut.ac.zw, edwinrupi@gmail.com

How to cite this paper: Chimwanda, P., & Rupi, E. (2026). Revisiting Logistic Regression for Diabetes Prediction: An Interpretable and High-Performance Analysis Using a Public Health Dataset. *Voice of the Publisher, 12,* 205-216.
https://doi.org/10.4236/vp.2026.122014

Received: February 4, 2026
Accepted: May 26, 2026
Published: May 29, 2026

Abstract

Logistic regression remains one of the most widely applied statistical techniques for modelling binary health outcomes due to its simplicity, interpretability, and robust theoretical foundation. Despite the growing popularity of complex machine learning algorithms in disease prediction, the continued relevance of traditional statistical models warrants empirical reassessment. This study re-examines the effectiveness of logistic regression for predicting diabetes status using a publicly available Kaggle dataset comprising demographic and symptomatic variables. A quantitative, cross-sectional research design was adopted, with diabetes status coded as a binary outcome. Model assumptions were assessed through multicollinearity diagnostics, while overall performance was evaluated using deviance statistics, Akaike Information Criterion (AIC), pseudo-R^2, and classification measures derived from a confusion matrix. The fitted logistic regression model demonstrated excellent predictive performance, achieving an overall accuracy of 93.3%, sensitivity of 94.4%, and specificity of 91.5%. Multicollinearity was not found to be a concern, supporting the stability of coefficient estimates. Several predictors, including polyuria, polydipsia, polyphagia, genital thrush, irritability, partial paresis, age, and gender, emerged as statistically significant determinants of diabetes status. The estimated odds ratios confirmed the strong clinical relevance of key symptomatic indicators, particularly polyuria and polydipsia. The findings highlight that logistic regression can rival more complex machine learning models in predictive accuracy while maintaining superior interpretability and lower computational demands. This study reinforces the continued value of logistic regression as a reliable and transparent tool for diabetes risk prediction, clinical decision support, and public

health screening initiatives.

Keywords

Logistic Regression, Diabetes Prediction, Binary Classification, Health Analytics, Model Interpretability, Confusion Matrix

1. Introduction

Logistic regression is a fundamental statistical modelling technique used to predict the probability of a binary outcome based on one or more predictor variables (Efunniyi et al., 2024). Unlike linear regression, which models continuous outcomes, logistic regression is specifically designed for categorical responses, typically coded as the occurrence or non-occurrence of an event. This makes it particularly suitable for applications where outcomes are dichotomous, such as success versus failure, presence versus absence of a condition, or survival versus death.

In the health sciences, logistic regression is one of the four principal multivariable statistical models, alongside linear regression, proportional hazards regression, and discriminant analysis (Tetrault et al., 2008). Among these, logistic regression is the most widely used due to its flexibility, interpretability, and suitability for binary outcomes. Its ability to estimate odds ratios allows researchers to quantify the strength and direction of associations between predictors and outcomes, thereby supporting evidence-based decision-making.

Linear regression is not suitable for these binary classification tasks primarily because its predictions are not constrained to the [0, 1] range required for probabilities. In addition, linear regression is strongly impacted by outliers, leading to inaccurate predictions and compromising the model's reliability. Logistic regression addresses these issues by modelling the probability that a given input belongs to a particular class. It uses the logistic function (sigmoid function) to ensure that the predicted probabilities are always between 0 and 1.

The model is extensively applied in medical and public health research to assess risk factors, predict health outcomes, and inform preventive interventions (Olowe et al., 2024; Ekpobimi et al., 2024). For example, logistic regression has been employed to estimate the likelihood of developing chronic conditions based on demographic, behavioural, and clinical variables. The odds of an event occurring are defined as the ratio of the probability that the event happens to the probability that it does not happen, providing an intuitive interpretation of model results (Adeniran et al., 2024).

Given the growing availability of large health datasets and the increasing need for accurate risk prediction, logistic regression remains a foundational analytical tool. This study employs logistic regression on a publicly available dataset to evaluate predictive performance and assess classification accuracy using a confusion

matrix.

The remaining part of the article consists of Literature Review, Methodology, Datta analysis, Discussion of findings, Conclusion and Future Research.

2. Literature Review

Logistic regression has a long history of application in health research due to its robustness in modelling binary and categorical outcomes. It enables researchers to examine the relationship between multiple independent variables and a dichotomous dependent variable while controlling for confounding factors (Kassem et al., 2022). Its interpretability, particularly through odds ratios, makes it especially valuable in clinical and epidemiological studies.

Despite its advantages, logistic regression faces several methodological challenges. One common issue is multicollinearity, which occurs when predictor variables are highly correlated with one another. Multicollinearity can inflate standard errors, reduce statistical power, and complicate the interpretation of model coefficients (Kassem et al., 2024). Another major challenge is overfitting, where the model becomes overly complex and captures noise rather than the underlying data structure, resulting in poor generalisation to new datasets (Ibikunle et al., 2024).

Logistic regression is often compared with more complex machine learning methods in predictive modelling tasks. The technique was among the seven techniques that were applied by Amilo et al. (2025) to predict diabetes risk in women. The other methods were Decision Tree, Support Vector Machine (SVM), Random Forest, Bagged Trees, Naive Bayes and XGBoost.

A table was presented comparing seven diabetes prediction models based on 10-fold cross-validation, evaluated across seven performance dimensions, namely accuracy, accuracy variability, precision, recall, F1-score, Area Under the Curve (AUC), and computation time. XGBoost demonstrated to be the strongest discriminative performance, achieving the highest AUC (0.839) and the highest recall (0.783), which indicates superior ability to correctly identify diabetic cases across varying decision thresholds, consistent with the AUC formulation described above the table. However, this performance came at the cost of the longest computation time (0.464 s), making it less suitable for time-critical or resource-constrained settings.

Logistic Regression provided respectable and stable performance with moderate accuracy (0.740) and AUC (0.832), while remaining computationally efficient (0.016 s). Despite not being the top performer, its interpretability makes it valuable.

Osibanjo et al. (2015) used Logistic Regression in determining the factors that contribute gaining admission at Faculty of business, University of Lagos, Nigeria. Admission was the dependent variable, with values admitted and not admitted. A sample of 395 students was taken from those who wanted to join the university from the foundation programme. The variables of interest included department,

gender, age, type of secondary school attended, difference between year of completion of secondary school and year of admission into the programme, total grade points in English, Mathematics and Economics at ordinary level, entrance examination score, mode of school fees payment at first registration, sponsor, first semester grade point average, and area of residence during the programme that concerned the graduates admitted into the programme during these aforementioned sessions and completed the programme for one academic session.

The results show that type of secondary school attended, mode of school fees payment at first registration, sponsor and first semester grade point average contribute significantly to the chance of gaining admission to the degree programme of the institution.

Mushtaq et al. (2022) introduced a highly accurate method for predicting diabetes based on a hybrid machine learning framework. Their approach utilizes a voting classifier that integrates multiple models, reaching an accuracy of 81.7%. Developed for the PIMA Indian Diabetes Dataset, the method incorporates comprehensive preprocessing steps, including outlier detection, missing value imputation, and oversampling, alongside ensemble learning techniques to improve predictive performance.

Similarly, Zafar et al. (2023) reviewed various diabetes prediction methods applied by different researchers. They highlighted a common limitation in prior studies—namely, the removal of rows containing missing values—which can negatively impact both parameter estimation and predictive accuracy. Overall, the method proposed by Mushtaq et al. is regarded as one of the most effective approaches for diabetes prediction.

Saoji et al. (2025) identifies significant situations that call for the use of Logistic Regression in Healthcare. Cardiovascular Disease Prediction, Cancer Prognosis and Recurrence, Infectious Disease Outcomes' Neurological and Respiratory Conditions and Cesarean Section Risk Prediction are the areas referenced in the article. A table with four columns showing the condition, key predictors, model purpose and reference. Prediction was found to be the most common purpose of the model, followed by identification of risk factors. It can be established from these findings that Logistic Regression is a powerful prediction tool.

Comparative studies have highlighted both the strengths and limitations of logistic regression relative to these methods. Logistic regression performs well in terms of recall and is effective for linearly separable data, although it may struggle with complex nonlinear relationships. Nonetheless, its simplicity, transparency, and lower computational cost continue to make it a preferred baseline model in many predictive studies.

While advanced machine learning models often achieve higher predictive accuracy, they frequently sacrifice interpretability and require extensive computational resources. There remains a need to reassess the effectiveness of traditional statistical models, such as logistic regression, using modern, publicly available datasets. In particular, limited studies focus on evaluating logistic regression perfor-

mance on Kaggle-based diabetes datasets while addressing common methodological challenges such as multicollinearity and overfitting.

This study addresses this gap by re-examining logistic regression as a predictive tool for diabetes, emphasizing model interpretability, performance evaluation, and comparative analysis. By doing so, it contributes to the ongoing discourse on balancing accuracy and interpretability in health predictive analytics.

3. Methodology

3.1. Research Design

This study adopted a quantitative, cross-sectional research design to re-examine the effectiveness of logistic regression in predicting diabetes status. Logistic regression was selected due to its suitability for modelling binary outcomes and its widespread application in health research. The analysis focused on estimating the probability of diabetes occurrence based on demographic and symptomatic predictors.

3.2. Data Source and Description

The dataset used in this study was obtained from a publicly accessible data repository, Kaggle, which is widely used for data science research and benchmarking. It is entitled "Diabetes Risk Prediction." It is version 1, has 520 rows, with no missing values, and 17 variables. Age is the only numeric variable with minimum 16, maximum 90, mean 48.03 and standard deviation 12.15. The rest of the variables are dichotomous with yes or no responses. It can be accessed at https://www.kaggle.com/datasets/rcratos/diabetes-risk-prediction. The dataset comprises medical, demographic, and symptomatic variables commonly associated with diabetes. The dependent variable, class, indicates diabetes status, categorized as Positive (diabetic) or Negative (non-diabetic).

Independent variables included age, gender, and a range of clinical symptoms such as polyuria, polydipsia, polyphagia, sudden weight loss, weakness, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity. Most predictors were binary categorical variables coded as *Yes* or *No*, while age was treated as a continuous variable.

3.3. Data Preparation and Coding

Prior to analysis, the dataset was screened for completeness, consistency, and coding accuracy. Categorical variables were dummy-coded, with appropriate reference categories specified. The outcome variable was coded such that *Positive* diabetes status represented the event of interest. Data preprocessing and model estimation were conducted using Jamovi, ensuring consistency in parameter estimation and diagnostic testing.

The outcome variable was coded as 1 = Positive (diabetic) and 0 = Negative (non-diabetic). For all dichotomous variables with yes/no responses, no was coded 0 and yes was coded 1. The reference category was 0 in each of these cases.

For gender, Female was used as the reference category (0), and Male was coded as 1. Thus, the odds ratio for gender (OR = 0.0129) reflects the odds of diabetes for males relative to females. This means that the odds of a male person being diabetic are 0.0129 times those of women.

3.4. Model Specification

A binary logistic regression model was fitted to estimate the log odds of having diabetes as a function of the selected predictors. The general form of the model is expressed as:

The logistic function is defined as: $\mathrm{P}\left(\mathrm{Y}=1\middle|\mathrm{X}\right)=\frac{1}{1+\mathrm{e}^{-z}}$, where

$$z=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k=\log\left(\frac{p}{1-p}\right)=\log it\left(p\right),where\ 0<p<1.$$

The parameter p represents the probability that the event occurs. In this case it is the probability of a positive diabetes diagnosis. β_0 is the intercept and β_k are regression coefficients associated with predictor variables X_k.

3.5. Model Validation Strategy

To assess generalization performance, the holdout method was used. The dataset was randomly partitioned into two equal subsets, with one used for training and the other for testing. Performance metrics for the two are shown in Table 1.

Table 1. Train and test set performance metrics.

Accuracy	Specificity	Sensitivity	AUC
0.958	0.918	0.967	0.992
0.942	0.947	0.936	0.988

The top row is from training while the bottom row is from testing. The train and test metrics are very close across all measures. The small decrease in accuracy and AUC, combined with the sensitivity-specificity trade-off, suggest good generalization. The area under the curve, which is 0.99 suggests that the model has excellent predictive performance.

4. Data Analysis and Results

4.1. Model Fit and Collinearity

Multicollinearity was assessed using the Variance Inflation Factor (VIF) and tolerance statistics, Table 2. All VIF values ranged between 1.27 and 2.92, and tolerance values exceeded 0.30, indicating that multicollinearity was not a significant concern. These results confirm the suitability of the predictors for inclusion in the logistic regression model.

Model fit was evaluated using deviance statistics, the Akaike Information Criterion (AIC), and the pseudo-coefficient of determination (R^2N). Classification

performance was assessed using a confusion matrix, sensitivity, specificity, and overall accuracy at a cut-off probability of 0.5. Based on the provided metrics, Table 3, the logistic regression model includes 17 estimated parameters (typically 1 intercept and 16 predictors), and demonstrates a strong fit to the data, as the residual deviance of 172 is substantially lower than the 520 – 17 = 503 degrees of freedom.

Table 2. Collinearity statistics for assumption checks.

	VIF	Tolerance
Age	2.30	0.434
Gender	1.98	0.505
Polyuria	2.37	0.422
Polydipsia	1.89	0.529
Sudden weight loss	1.55	0.643
Weakness	1.82	0.550
Polyphagia	1.53	0.655
Genital thrush	1.60	0.624
Visual blurring	2.56	0.391
Itching	2.92	0.342
Irritability	1.43	0.701
Delayed healing	1.86	0.539
Partial paresis	1.48	0.678
Muscle stiffness	1.75	0.572
Alopecia	2.36	0.424
Obesity	1.27	0.786

Table 3. Model fit and overall performance.

Model	Deviance	AIC	R^2_N
1	172	206	0.860

The Nagelkerke's pseudo-R^2 measure in Table 3 has a value of 0.86. It is an additional statistic that can be used to evaluate the usefulness of a model. This statistic ranges from 0 to 1 and is often preferred because it allows for easier interpretation. R^2 indicates the extent to which the explanatory variables contribute to predicting the response variable and is best interpreted as a measure of effect size. A Nagelkerke R^2 value of 0.860 suggests that the model has strong predictive usefulness.

The confusion matrix, Table 4, shows excellent predictive performance. The model correctly classified 91.5% of non-diabetic cases and 94.4% of diabetic cases.

From Table 5, overall model accuracy was 93.3%, with a sensitivity of 94.4% and specificity of 91.5%, suggesting that the model was highly effective in identi-

fying both diabetic and non-diabetic individuals.

Table 4. Confusion matrix

Observed	Predicted		% Correct
	Negative	Positive	
Negative	183	17	91.5
Positive	18	302	94.4

Note. The cut-off value is set to 0.5.

Table 5. Predictive measures.

Accuracy	Specificity	Sensitivity
0.933	0.915	0.944

Note. The cut-off value is set to 0.5.

4.2. Interpretation of Model Coefficients

Table 6. Model coefficients - class.

Predictor	Estimate	SE	Z	*p*	Odds ratio	95% confidence interval	
						Lower	Upper
Intercept	2.7466	1.0755	2.554	0.011	15.5895	1.89404	128.3140
Age	−0.0512	0.0254	−2.017	0.044	0.9501	0.90404	0.9985
Gender	−4.3512	0.5982	−7.274	<0.001	0.0129	0.00399	0.0416
Polyuria	4.4395	0.7053	6.295	<0.001	84.7359	21.26884	337.5913
Polydipsia	5.0704	0.8289	6.117	<0.001	159.2446	31.37076	808.3590
Sudden weight loss	0.1903	0.5477	0.348	0.728	1.2096	0.41352	3.5385
Weakness	0.8171	0.5368	1.522	0.128	2.2638	0.79053	6.4830
Polyphagia	1.1938	0.5335	2.238	0.025	3.2995	1.15962	9.3880
Genital thrush	1.8637	0.5533	3.368	<0.001	6.4472	2.17969	19.0701
Visual blurring	0.9159	0.6512	1.406	0.160	2.4990	0.69735	8.9551
Itching	−2.8029	0.6727	−4.167	<0.001	0.0606	0.01622	0.2266
Irritability	2.3407	0.5905	3.964	<0.001	10.3888	3.26519	33.0539
Delayed healing	−0.3916	0.5500	−0.712	0.476	0.6760	0.23000	1.9866
Partial paresis	1.1593	0.5248	2.209	0.027	3.1877	1.13961	8.9167
Muscle stiffness	−0.7288	0.5802	−1.256	0.209	0.4825	0.15475	1.5044
Alopecia	0.1504	0.6201	0.242	0.808	1.1623	0.34473	3.9185
Obesity	−0.2890	0.5443	−0.531	0.595	0.7490	0.25771	2.1768

Note. Estimates represent the log odds of "class = 1" vs. "class = 0".

Several predictors emerged as statistically significant determinants of diabetes sta-

tus, Table 6. Age was significantly negatively associated with diabetes, indicating a slight decrease in odds with increasing age when controlling for symptoms. Gender showed a coefficient of −4.3512 for male, with OR = 0.0129 and $p < 0.001$. This means that when modeling diabetes status (1 = Yes, 0 = No) using a gender indicator (1 = Male, 0 = Female), a coefficient of −4.3512 for the male variable indicates that males have 4.3512 lower log-odds of being diabetic compared to females. Exponentiating this coefficient gives an odds ratio of $e^{-4.3512} = 0.0129$. This means that, holding all other variables constant, the odds of being diabetic for males are 0.0129 times the odds for females. The coefficient of age is −0.0512 which means that an increase of one year in age decreases the log-odds by −0.0512.

Symptoms such as polyuria (OR = 84.74, $p < 0.001$) and polydipsia (OR = 159.24, $p < 0.001$) exhibited extremely strong positive associations with diabetes, which may reflect near-separation inherent in symptom-based screening datasets, and future studies should consider penalized logistic regression to mitigate potential overestimation. Polyphagia (OR = 3.30, $p = 0.025$), genital thrush (OR = 6.45, $p < 0.001$), irritability (OR = 10.39, $p < 0.001$), and partial paresis (OR = 3.19, $p = 0.027$) were also significant predictors.

Conversely, itching showed a significant negative association with diabetes (OR = 0.06, $p < 0.001$), suggesting a complex or context-dependent relationship. Other variables, including sudden weight loss, weakness, visual blurring, delayed healing, muscle stiffness, alopecia, and obesity, were not statistically significant within the multivariable framework.

4.3. Predictions

This example demonstrates how predicted probabilities depend strongly on the baseline covariate pattern and should not be interpreted as a clinical rule. The model

$$\mathrm{P}\left(\mathrm{Y}=1\middle|\mathrm{X}\right)=\frac{1}{1+\mathrm{e}^{-(2.75-0.05A-4.35G+4.44P_1+5.07P_2+1.19P_1 1.86\mathrm{GT}-2.80\mathrm{IC}+2.34\mathrm{IR})+1.1593\mathrm{PP}}}$$

was used in the estimation of the probabilities. To illustrate predicted probabilities, consider a 45-year-old female reporting irritability only (all other symptoms = No). Substituting these values into the linear predictor yields:

$$\left(\mathrm{z}=2.75-0.05\times 45-4.35\times 0+2.34\times 1\right)=2.84$$

The corresponding predicted probability is …

$$\frac{1}{1+\mathrm{e}^{-(2.84)}}=0.945.$$

This example demonstrates how predicted probabilities depend strongly on the baseline covariate pattern and should not be interpreted as a clinical rule.

5. Discussion of Findings

The findings reaffirm the robustness of logistic regression as a predictive tool for

diabetes classification, even in the era of advanced machine learning algorithms. The model demonstrated excellent classification accuracy, sensitivity, and specificity, exceeding the performance reported in some prior studies using classical statistical methods and rivalling that of more complex models.

The strong effects observed for polyuria and polydipsia align with established clinical knowledge, reinforcing their role as hallmark symptoms of diabetes. The statistical significance of polyphagia, genital thrush, irritability, and partial paresis further highlights the importance of symptomatic indicators in predictive modelling.

While some variables traditionally associated with diabetes, such as obesity and weight loss, were not significant in this model, this may reflect interrelationships among symptoms or dataset-specific characteristics. Importantly, multicollinearity diagnostics confirmed that correlated predictors did not distort coefficient estimates, strengthening confidence in the model results.

Compared with machine learning approaches reported in the literature, logistic regression demonstrated competitive performance while maintaining superior interpretability. This supports previous findings that logistic regression performs well for linearly separable health data and remains highly valuable for clinical decision-making and policy formulation.

6. Conclusion

This study revisited the application of logistic regression for diabetes prediction using a Kaggle dataset and Jamovi analytical tools. The results demonstrate that logistic regression remains a powerful, interpretable, and statistically sound method for predicting diabetes status. With an overall accuracy of 93.3% and strong sensitivity and specificity, the model effectively identified individuals at risk of diabetes.

The findings confirm that classical statistical models continue to hold relevance in modern health analytics, particularly when transparency, interpretability, and clinical applicability are prioritized. Logistic regression therefore remains a dependable approach for disease risk prediction and early screening initiatives.

7. Future Research

The study recommends that logistic regression models be integrated into routine clinical diabetes screening, especially in resource-constrained settings where model simplicity and interpretability are critical for effective decision-making. Future research should also consider hybrid modelling approaches that combine logistic regression with advanced machine-learning techniques in order to enhance predictive accuracy while retaining transparency. In addition, the use of larger and more diverse clinical datasets is encouraged to improve the stability, robustness, and generalizability of model estimates across different populations. Model performance may be further strengthened through feature engineering and nonlinear extensions, such as the inclusion of interaction terms or the application

of regularized logistic regression to better capture complex relationships among symptoms. Finally, policymakers and public health practitioners are advised to adopt interpretable models like logistic regression to inform early detection strategies, guide targeted interventions, and support evidence-based public health planning.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Adeniran, I. A., Abhulimen, A. O., Obiki-Osafiele, A. N., Osundare, O. S., Agu, E. E., & Efunniyi, C. P. (2024). Strategic Risk Management in Financial Institutions: Ensuring Robust Regulatory Compliance. *Finance & Accounting Research Journal, 6,* 1582-1596. https://doi.org/10.51594/farj.v6i8.1508

Amilo, D., Sadri, K., Hincal, E., Farman, M., Nisar, K. S., & Hafez, M. (2025). An Integrated Machine Learning and Fractional Calculus Approach to Predicting Diabetes Risk in Women. *Healthcare Analytics, 8,* Article ID: 100402. https://doi.org/10.1016/j.health.2025.100402

Efunniyi, C. P., Abhulimen, A. O., Obiki-Osafiele, A. N., Osundare, O. S., Agu, E. E., & Adeniran, I. A. (2024). Strengthening Corporate Governance and Financial Compliance: Enhancing Accountability and Transparency. *Finance & Accounting Research Journal, 6,* 1597-1616. https://doi.org/10.51594/farj.v6i8.1509

Ekpobimi, H. O., Kandekere, R. C., & Fasanmade, A. A. (2024). Software Entrepreneurship in the Digital Age: Leveraging Front-End Innovations to Drive Business Growth. *International Journal of Engineering Research and Development, 20.*

Ibikunle, O. E., Usuemerai, P. A., Abass, L. A., Alemede, V., Nwankwo, E. I., & Mbata, A. O. (2024). Artificial Intelligence in Healthcare Forecasting: Enhancing Market Strategy with Predictive Analytics. *International Journal of Applied Research in Social Sciences, 6,* 2409-2446.

Kassem, R. G., Mbata, A. O., Usuemerai, P. A., Abass, L. A., & Ogbewele, E. G. (2022). Digital Transformation in Pharmacy Marketing: Integrating AI and Machine Learning for Optimized Drug Promotion and Distribution. *World Journal of Advanced Research and Reviews, 15,* 749-762. https://doi.org/10.30574/wjarr.2022.15.2.0792

Kassem, R. G., Mbata, A. O., Usuemerai, P. A., Abass, L. A., & Ogbewele, E. G. (2024). Pharmacy Marketing for Public Health Impact: Promoting Preventive Care and Health Literacy through Strategic Campaigns. *World Journal of Advanced Research and Reviews, 18,* 1406-1418. https://doi.org/10.30574/wjarr.2023.18.2.0982

Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., & Husnain, M. (2022). Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques. *Mobile Information Systems, 2022,* Article ID: 6521532. https://doi.org/10.1155/2022/6521532

Olowe, K. J., Edoh, N. L., Zouo, S. J. C., & Olamijuwon, J. (2024). Comprehensive Review of Logistic Regression Techniques in Predicting Health Outcomes and Trends. *World Journal of Advanced Pharmaceutical and Life Sciences, 7,* 16-26.

Osibanjo, F. S., Olalude, G. A., Akintunde, M. O., & Ajala, A. G. (2015). Application of Logistic Regression Model to Admission Decision of Foundation Programme at University of Lagos. *International Journal of Mathematics and Statistics Studies, 3,* 27-41.

Saoji, P., Madireddy, L., & Saoji, A. (2025). Logistic Regression in Healthcare: Predictive Modelling for Enhanced Clinical Decision-Making. *Journal of Neonatal Surgery, 14,* 765-769.

Tetrault, J. M., Sauler, M., Wells, C. K., & Concato, J. (2008). Reporting of Multivariable Methods in the Medical Literature. *Journal of Investigative Medicine, 56,* 954-957. https://doi.org/10.2310/jim.0b013e31818914ff

Zafar, A., Tahir, A., & Asgher, U. (2023). A Review of Diverse Diabetic Prediction Models: A Literature Study. *TIERS Information Technology Journal, 4,* 150-164. https://doi.org/10.38043/tiers.v4i2.3617