Orginal Article
Open Access
Iranian Journal of Diabetes and Obesity (IJDO)

Mathematical Foundations of Diabetes Forecasting Studies: A Comparative Analysis of Statistics and ML Models

Alireza Pakgohar*[1]
Department of Statistics, Payame Noor University, Tehran, Iran.

Abstract

This study provides a systematic comparison of the mathematical properties, strengths, and limitations of traditional statistical methods and machine learning models in diabetes forecasting. While classical approaches like logistic regression and ANOVA offer interpretability and simplicity, their reliance on linear assumptions and sensitivity to heteroscedasticity limit their utility in modeling complex, nonlinear relationships inherent in diabetes data. In contrast, machine learning techniques-including neural networks, random forests, and gradient boosting-excel in capturing high-dimensional interactions and nonlinear dynamics, achieving superior predictive accuracy. However, these gains come at the cost of computational complexity, black box interpretability challenges, and ethical concerns around algorithmic bias. Through a detailed analysis of mathematical frameworks (e.g., activation functions, regularization, ensemble methods), we demonstrate how hybrid approaches integrating explainable AI (XAI) can bridge the gap between statistical rigor and clinical usability. Our findings highlight the critical trade-offs between model interpretability, predictive power, and scalability, offering actionable insights for optimizing diabetes risk prediction in precision medicine.

Keywords: Statistical modeling, Machine learning, Predictive analytics, Sample size, Big data

QR Code:

Citation: Pakgohar A. Mathematical Foundations of Diabetes Forecasting Studies: A Comparative Analysis of Statistics and ML Models. IJDO 2026; 18 (2) :113-124

URL: http://ijdo.ssu.ac.ir/article-1-1038-en.html

10.18502/ijdo.v18i2.21648

Article info:
Received: 22 March 2025
Accepted: 17 May 2026
Published in June 2026

Corresponding Author:
Alireza Pakgohar, Department of Statistics, Payame Noor University, Tehran, Iran.

Tel: (98) 913 318 0781
Email: a_pakgohar@pnu.ac.ir

Orcid ID: 0000-0003-2512-1581

[DOI: 10.18502/ijdo.v18i2.21648]

Introduction

Classical statistical methods, which employ mathematical formulas and assumptions for data analysis and prediction, though effective in identifying basic patterns, often face limitations in capturing complex, nonlinear, or multidimensional relationships-such as modeling intricate patterns like nonlinear interactions among insulin resistance, metabolic pathways, and comorbidities in diabetes and exhibit restricted capability in encapsulating real world complexities (1,2).

In contrast to classical statistical approaches, which rely on rigid assumptions and often oversimplify nonlinear relationships, machine learning algorithms are capable of efficiently processing large datasets, adapting in real time, and uncovering hidden patterns without requiring predefined assumptions.

For instance, in fields such as diabetes research where disease onset results from complex interactions among genetic, environmental, and lifestyle factors traditional tools like modern statistical science encompasses generalized linear models (GLMs), generalized additive models (GAMs) for nonlinear relationships, and quantile regression methods for addressing heteroscedasticity. These methods struggle with nonlinear dependencies and rely on assumptions such as linearity, normality, and data homogeneity, which do not always hold in real-world data. In contrast, machine learning excels at modeling multifactorial systems and provides deeper insights into causal relationships and correlations that would otherwise remain obscured (3).

For example, identifying shared disease genes in metabolic syndrome and cardiovascular diseases requires integrated network analyses a task beyond the capabilities of classical statistics. Novel alternatives such as neural networks and decision trees (4-5) address these gaps more effectively by modeling nonlinear dynamics and multidimensional interactions. These methods detect latent patterns in complex datasets, such as predicting therapeutic responses in heterogeneous populations (6-7).

However, uncertainties regarding transparency and interpretability may hinder the adoption of these tools in clinical settings. This implies that the implementation of machine learning models must ensure that physicians and healthcare professionals can comprehend their outputs and derive actionable clinical decisions (8-10).

This research focuses on several key areas to provide a comprehensive understanding of statistical and machine learning predictive models, with an emphasis on outcomes obtained in the context of diabetes. To this end, an effort is made to examine traditional statistical models and evaluate their limitations. Subsequently, advanced machine learning algorithms and their mathematical properties will be analyzed.

Challenges and limitations in statistical approaches

Tools such as statistical significance and effect size not only form the basis of hypothesis testing but also enhance the interpretability and evaluation of machine learning models. However, these methods have limitations. Model misspecification, violation of assumptions, and sensitivity to outliers can lead to biased results and reduce the reliability of findings. For instance, heteroscedasticity unequal variance in residuals presents a significant challenge, especially in complex medical datasets. To maintain methodological integrity, it is essential to address these issues through robust model selection, assumption checking, and validation techniques.

Heteroscedasticity testing and correction

For example, heteroscedasticity-the presence of non-constant variance in model residuals-poses a significant challenge in regression-based studies within the field of diabetes. This

[DOI: 10.18502/ijdo.v18i2.21648]

can lead to biased parameter estimates and an increase in Type I/Type II errors, thereby undermining clinical inferences (10). Traditional remedies, such as weighted least squares or data transformations, often fail to provide actionable insights for clinicians due to their obscure mathematical formulations and limited interpretability.

Sample size and P-value

The reliance of traditional statistical models on *P*-values, along with the inherent dependence of these values on sample size, imposes limitations on classical methods, particularly in human clinical trials and animal studies where small sample sizes often lead to reduced statistical power. To address this, we propose integrating Machine Learning algorithms not as a replacement, but as a complementary analytical framework. These methods can detect complex, non-linear patterns and enhance predictive robustness in small datasets, thereby mitigating the limitations associated with exclusive reliance on *P*-value based inference.

Conversely, increasing sample sizes raises ethical concerns in medical research. Furthermore, in studies involving very large populations, the sensitivity of *P*-values causes them to converge towards zero, rendering conventional hypothesis testing trivial. To address these limitations, contemporary methodology emphasizes estimation-based inference and resampling techniques, such as bootstrapping. These alternatives prioritize the assessment of effect sizes and confidence intervals, thereby providing a more robust framework for statistical inference than reliance on p-values alone. For more details see (13-16).

Machine learning models

One of the most important machine learning-based models is neural network models. Neural networks have shown remarkable performance in complex prediction tasks. One of the main applications of neural network models is medical prediction which has shown very good performance and is being developed day by day (17). For example, an artificial neural network-based model achieved 96.8% accuracy in predicting diabetes, which surpassed traditional methods (18-19). It should be noted that we are talking about the accuracy measure and not the precision, however, the number 97% accuracy indicates the high reliability of the model in handling complex data sets.

For explicit mentions of precision, a deep belief network (DBN) method demonstrated 82% precision in diabetes detection, though this is lower than the accuracy reported in (20). The discrepancy underscores the importance of evaluating multiple metrics (e.g., accuracy, sensitivity, specificity) when assessing model performance in clinical contexts. Similarity, Zabbah et al, (21) indicate that the ANFIS classifier has significant potential in accurately detecting diabetes, highlighting its applicability in medical diagnosis systems (Accuracy more than 80%).

Note that While (13-14) report 96.8% accuracy (overall correctness), (15) highlights 82% precision (minimizing false positives), underscoring the need for multi-metric evaluations. While accuracy measures overall correctness, precision reflects the model's ability to avoid false positives. The lower precision in (15) suggests trade-offs between sensitivity and specificity.

These models have further improved predictive capabilities by optimizing the training algorithms. These models are particularly useful in complex cases where nonlinear factors between genes (such as metabolic markers and epigenetic factors) are critical (22).

Challenges and limitations in machine learning

Machine learning models offer strong predictive capabilities in healthcare, improving diagnosis, treatment, and patient outcomes. However, their adoption is limited by key challenges, including high data demands, lack of interpretability due to their "black box"

[DOI: 10.18502/ijdo.v18i2.21648]

nature, and ethical concerns. Addressing these issues is essential to develop fair, transparent, and trustworthy ML-driven healthcare solutions (23).

A principal impediment in deploying machine learning (ML) within healthcare is the models' dependence on extensive and heterogeneous datasets. While such datasets encompassing patient demographics, clinical history, diagnostics, and outcomes are foundational for robust predictive performance, their complexity introduces substantial analytic challenges. Data incompleteness, inconsistency, and variability can propagate noise and systematic bias, particularly when training sets lack demographic diversity. This limitation exacerbates the risk of algorithmic bias, wherein predictive accuracy diminishes for underrepresented groups, potentially leading to inequitable healthcare decisions and outcomes. Further details; please refer to (24-25).

Another major challenge is the "black box" problem inherent in many ML algorithms, particularly in complex models like deep neural networks (26). These models often lack transparency, making it difficult for healthcare providers to understand or explain the reasoning behind their predictions. Topol (2019) argues healthcare professionals are unlikely to embrace AI systems that fail to offer clear explanations for their predictions, since trust and accountability are crucial in the medical field (26). This opacity poses a significant barrier to the adoption of AI in clinical settings, as clinicians are understandably hesitant to trust systems whose decision-making processes they cannot comprehend. Without interpretability, it becomes challenging to validate the accuracy and fairness of AI-driven recommendations, potentially compromising patient safety and trust in the technology.

In conclusion, while machine learning offers transformative potential for healthcare, addressing its challenges is essential to ensure that AI solutions are equitable, interpretable, and trustworthy. By focusing on data quality, transparency, and ethical considerations, the healthcare industry can harness the power of AI to improve patient outcomes while minimizing potential risks. This will require interdisciplinary collaboration, robust governance frameworks, and a commitment to global equity in AI development and deployment.

Mathematical properties of models in diabetes forecasting: foundations and advanced concepts

Cross-validation is a cornerstone statistical technique for robust model evaluation. K-fold cross-validation improves reliability by partitioning data into multiple subsets and averaging performance across folds. Incorporating stratified sampling preserves class proportions, crucial for imbalanced clinical datasets. Advanced methods such as nested and repeated cross-validation further enhance rigor by isolating model selection from assessment and quantifying stability, thereby ensuring trustworthy evaluation of complex models with extensive hyper-parameter tuning (28).

Statistical robustness is established through comprehensive diagnostic procedures including detailed residual analysis, systematic outlier detection and influence assessment, multicollinearity evaluation using variance inflation factors, and thorough examination of model calibration and discrimination metrics. These methodological safeguards collectively ensure that predictive models maintain their validity across diverse data conditions and resist undue influence from anomalous observations or problematic data structures, ultimately enhancing the reliability and generalizability of analytical findings in research and clinical applications (12).

Data patterns and relationships

Logistic regression emerges as a powerful probabilistic modeling technique for diabetes

[DOI: 10.18502/ijdo.v18i2.21648]

risk prediction, utilizing a sophisticated mathematical framework defined by the probability function (Equation 1)

$$P(Y=1)=1/(1+e^{-z}), \quad (1)$$

Where z represents a linear combination of weighted predictive features. The model comprehensively integrates diverse risk determinants including anthropometric indicators (BMI, body composition), metabolic parameters (blood glucose, insulin resistance markers), demographic characteristics, genetic predispositions, and lifestyle factors such as physical activity and dietary patterns. By leveraging both classical and advanced Bayesian statistical approaches, the methodology enables nuanced risk stratification through sophisticated computational techniques like Probabilistic Model Checking (PMC) and Markov Chain Monte Carlo sampling (30). The model's strength lies in its ability to synthesize complex, multidimensional data sources, transforming heterogeneous information into interpretable probabilistic risk estimates. Advanced implementations focus on non-linear feature interactions, dynamic risk profiling, and uncertainty quantification, allowing for personalized and adaptive predictive insights. Validation strategies encompass rigorous statistical assessments including cross-validation, calibration curve analysis, and discrimination metrics, ensuring robust and reliable predictive performance. This approach represents a sophisticated methodological framework that transcends traditional risk assessment techniques, offering a comprehensive, data-driven approach to understanding and predicting diabetes risk across diverse population segments.

Cross-validation and regularization represent advanced statistical techniques designed to enhance predictive modeling's reliability and generalizability (29). The k-fold cross-validation approach systematically partitions datasets to comprehensively assess model performance, ensuring robust and reproducible results. Regularization methods, including L1 (Lasso), L2 (Ridge), and Elastic Net, strategically manage model complexity by preventing overfitting and addressing multi-collinearity through sophisticated penalty mechanisms (11,31). These techniques enable more nuanced feature selection and parameter estimation, particularly effective in handling complex predictor interactions such as age-BMI and glucose-medication relationships. By balancing model complexity, computational efficiency, and predictive accuracy, cross-validation and regularization provide a sophisticated framework for developing statistically rigorous and clinically meaningful predictive models across diverse analytical domains, ultimately improving the reliability and interpretability of statistical predictions.

Kernel methods, particularly the Radial Basis Function the RBF kernel (Equation 2) and polynomial kernel (Equation 3), are advanced mathematical techniques that transform linear, non-separable data into higher-dimensional feature spaces (32-33). By applying non-linear transformations represented by the RBF kernel (Equation 2

$$K(x,x') = exp\left(\frac{\|x-x'\|^2}{2\sigma^2}\right) \quad (2$$

The polynomial kernel captures higher-order feature interactions through non-linear transformations, enabling complex decision boundary modeling (Equation 3)

$$K_p(x,y) = \left(\gamma\|x-y\|+c\right)^2. \quad (3)$$

These methods enable complex decision boundary representation, capture intricate spatial relationships, and enhance predictive modeling across challenging datasets. The approach flexibly maps input vectors, measuring distance-based similarities and allowing sophisticated pattern recognition in machine learning applications, particularly in support vector machines and complex classification tasks.

Neural networks (NNs), inspired by the human brain, represent a fundamental machine learning paradigm, comprised of interconnected nodes (neurons) organized in layered structures designed for data processing, pattern recognition, and decision-

[DOI: 10.18502/ijdo.v18i2.21648]

making. These networks typically consist of an input layer, one or more hidden layers, and an output layer, with the connections between neurons characterized by associated weights that are iteratively adjusted during the learning process.

Neural networks, inspired by the human brain, consist of interconnected layers of neurons that transform input data through nonlinear activation functions (for example Rectified Linear Unit (ReLU) and sigmoid). The learning process involves minimizing a loss function using gradient descent, where the model iteratively adjusts weights to reduce prediction error. In contrast, random forests employ an ensemble approach, combining multiple decision trees trained on bootstrapped samples of the data. This ensemble method not only improves predictive accuracy but also provides insights into feature importance, making it a valuable tool for high-dimensional datasets (27).

A central element of NNs is the introduction of non-linearity through activation functions, such as (Equations 4 and 5)

$$\text{ReLU} = f(x) = \max(0, x) \quad (4)$$

$$Sigmoid : \sigma(x) = \frac{1}{1+e^{-x}}, \quad (5)$$

Enabling the modeling of complex relationships within the data. The fundamental computations within an NN involve the weighted sum of inputs (Equation 6)

$$\sum w_i x_i + \text{bias} \quad (6)$$

Also, the application of activation functions to transform these sums, with a basic linear model of a neuron expressed as (Equation 7)

$$y' = b + w_1 x_1 + w_2 x_2 + w_3 x_3 ,..., w_k x_k \quad (7)$$

Where y′ represents the output, b is the bias, w's are the weights, and x's are the inputs. Through this architecture and these key mathematical operations, NNs are capable of learning intricate patterns and making predictions across a wide array of applications.

A Comparative Analysis of Traditional Statistical Methods and Machine Learning Models

In this section, we have explored the mathematical properties of traditional statistical methods and their limitations in handling complex, high-dimensional data. In the following section, we will examine how machine learning models address these challenges and discuss their potential applications in diabetes forecasting.

Mathematical Foundations

Linear relationships in mathematics refer to relationships between variables that can be represented by a straight line on a graph. These relationships are often simpler to analyze and interpret because they follow a predictable pattern. Nonlinear relationships, on the other hand, do not follow a straight line and can be more complex to understand (28).

When working with linear relationships, certain assumptions are often made about the data being analyzed. For example, it is assumed that the relationship between variables is constant and that there is no interaction between them. These assumptions make it easier to build models and make predictions based on the data (28).

However, in real-world scenarios, many relationships are nonlinear and do not adhere to these assumptions. This can make modeling more challenging as the relationship between variables may be more difficult to define and predict accurately (32).

On the other hand, are better equipped to capture complex relationships and patterns in the data that may not be easily represented by a linear model (11). However, they can be more difficult to interpret and may require more computational resources to train and evaluate. In general, the choice between a linear or nonlinear approach will depend on the specific characteristics of the data and the goals of the analysis. It is important to carefully consider model complexity and

[DOI: 10.18502/ijdo.v18i2.21648]

interpretability when selecting an appropriate modeling approach.

Computational Efficiency

Computational efficiency is a critical factor in determining the effectiveness of a machine learning model. Training time refers to the amount of time it takes for a model to learn from the data and adjust its parameters accordingly. A model that can be trained quickly is advantageous, especially in scenarios where real-time decision-making is required.

Scalability is essential for managing large-scale datasets and intricate computations, allowing for efficient processing across distributed systems (27,33). A model that can scale effectively means that it can handle larger datasets and more complex problems without sacrificing performance. This is crucial for applications that require processing large amounts of data or dealing with high-dimensional input features.

Resource requirements also play a significant role in computational efficiency. Models that are resource-intensive may not be practical for deployment on devices with limited computing power or memory. Optimizing resource usage can help reduce costs and improve overall performance.

Interpretability

Interpretability refers to the ability to understand and explain how a model makes predictions. In machine learning, there is often a trade-off between predictive power and explain ability. Models that are highly complex and have high predictive power, such as deep neural networks, may be difficult to interpret and explain. For example in high-stakes fields such as healthcare and criminal justice, it is important to prioritize interpretable models over complex black-box models to maintain transparency and accountability (23). On the other hand, simpler models like decision trees or linear regression are more interpretable but may sacrifice some predictive power.

Finding the right balance between predictive power and explain ability is crucial in many applications, especially in fields where decisions have significant consequences, such as healthcare or finance. In these cases, it is important to not only make accurate predictions but also be able to understand why a model made a particular prediction. This can help build trust in the model's decisions and ensure that they are fair and unbiased. In healthcare, comprehending the reasons behind a model's prediction is just as crucial as the prediction itself, as it is essential for ensuring fairness, safety, and trust (25).

Researchers are constantly working on developing new techniques for improving the interpretability of complex models without sacrificing too much predictive power. Techniques such as feature importance analysis, local explanations, and model-agnostic methods aim to provide insights into how a model makes predictions and why certain decisions are made. By understanding the inner workings of complex models, researchers can ensure that the decisions made by these models are fair, transparent, and trustworthy (26). This is crucial in fields such as healthcare, finance, and criminal justice where the stakes are high and decisions have real-world consequences. Ultimately, the goal is to strike a balance between accuracy and interpretability so that complex models can be used effectively and ethically in a wide range of applications.

Handling High-Dimensional Data

One of the key challenges in handling high-dimensional data is the ability to effectively manage large datasets with many variables. High-dimensional data presents considerable challenges due to the curse of dimensionality, necessitating the use of advanced preprocessing and dimensionality reduction techniques (32). This requires not only robust storage and processing capabilities, but also advanced techniques for data cleaning, preprocessing, and analysis. With the increasing complexity and size of modern

[DOI: 10.18502/ijdo.v18i2.21648]

datasets, it is essential for researchers and data scientists to have the tools and skills necessary to navigate this high-dimensional space efficiently. Scalable frameworks such as Apache Spark facilitate the efficient processing and analysis of high-dimensional datasets by utilizing distributed computing (33).

Some common approaches to handling high-dimensional data include dimensionality reduction techniques such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE), which can help simplify complex datasets by capturing the most important features (30). Additionally, machine learning algorithms like random forests or support vector machines are often used to model relationships between variables in high-dimensional data.

Overall, the capacity to analyze high-dimensional data is essential for revealing hidden patterns and facilitating informed decision-making across various domains such as genomics, healthcare, and finance (32). By leveraging advanced tools and techniques for handling high-dimensional data, researchers can uncover hidden patterns, trends, and relationships that may not be apparent through traditional analysis methods. This can lead to more accurate predictions, improved decision-making processes, and ultimately, better outcomes in various industries. AI-powered analysis of high-dimensional data is revolutionizing healthcare by allowing for more precise predictions, tailored treatments, and enhanced patient outcomes (26).

Additionally, the ability to effectively manage large datasets can also help organizations identify potential risks and opportunities, optimize resource allocation, and drive innovation. In today's data-driven world, mastering the art of handling complex datasets is essential for staying competitive and driving success in a rapidly evolving landscape.

Robustness to Noise and Outliers

Robustness to noise and outliers is essential in ensuring that the model's performance is not significantly impacted by the presence of irrelevant or erroneous data points (32). Sensitivity to data quality issues refers to how well the model can handle variations in data quality, such as missing values, inconsistencies, or outliers. A robust model should be able to effectively filter out noise and identify outliers without being overly influenced by them. Efficient preprocessing methods, including outlier detection and data normalization, are essential for enhancing the robustness and generalizability of models (32). This requires careful preprocessing of the data and selection of appropriate algorithms that are resilient to such issues. Ultimately, a model's ability to maintain high performance despite variations in data quality is crucial for its reliability and generalizability. A model's capacity to generalize to new, unseen data depends on its resilience to fluctuations in data quality and its ability to manage noise effectively (27).

Table. Overview of the comparison between traditional statistical methods and ML models

Aspect	Traditional Statistical Methods	Machine Learning Models
Mathematical Foundation	Linear relationships, fixed model structures, parametric assumptions (e.g., normality, homoscedasticity).	Nonlinear relationships, flexible model structures, non-parametric or semi-parametric approaches.
Computational Efficiency	Low computational cost, suitable for small datasets.	High computational cost requires large datasets and significant computational resources.
Interpretability	High interpretability (e.g., coefficients in regression models).	Lower interpretability (e.g., "black box" nature of deep learning models).
Handling High-Dimensional Data	Limited ability to handle interactions among hundreds of variables.	Excels at capturing high-dimensional interactions (e.g, random forests, neural networks).
Robustness to Noise	Sensitive to outliers and violations of assumptions (e.g., heteroscedasticity).	More robust to noise and outliers, especially with regularization techniques.

[DOI: 10.18502/ijdo.v18i2.21648]

A synthesis of the methodological trade-offs and comparative characteristics discussed throughout this section is presented in Table, which provides a high-level overview of the divergence between traditional statistical methodologies and machine learning approaches in diabetes forecasting.

Discussion

The present study evaluated the mathematical properties of statistical models in the context of diabetes forecasting, specifically focusing on the trade-off between interpretability and predictive accuracy. A structured comparison reveals that while traditional statistical methods such as Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) offer robust frameworks for inference and understanding the marginal effects of predictors, they often lack the flexibility required to capture complex, high-dimensional non-linearities inherent in biological data. Our analysis indicates that reliance on classical assumptions, such as homoscedasticity and linearity, limits the utility of these models when analyzing heterogeneous patient profiles involving intricate interactions between genetic and metabolic factors.

In contrast, the application of advanced Machine Learning (ML) algorithms and Neural Networks demonstrated superior performance in modeling these non-linear dynamics. Unlike classical hypothesis testing, which is often constrained by sample size dependencies and the convergence of P-values in large datasets, ML approaches prioritize predictive robustness and generalizability. However, this gain in accuracy often comes at the cost of interpretability (the “black box” phenomenon). Therefore, rather than viewing these methodologies as mutually exclusive, our findings suggest a synergistic approach. By utilizing estimation-based inference and resampling techniques (e.g., bootstrapping) alongside Neural Networks, researchers can mitigate the limitations of traditional *P*-value driven methods while retaining the predictive power necessary for precision medicine. This comprehensive analytical framework ensures that the heterogeneity of variance is adequately addressed, providing a more reliable basis for clinical decision-making.

Conclusions

The mathematical properties of the models examined in this research underscore a pivotal shift in the paradigm of diabetes management. Our investigation confirms that while traditional statistical models provide a necessary foundation for causal inference, they are increasingly insufficient for the granular forecasting required in modern healthcare. The results substantiate the hypothesis that Machine Learning and Neural Network models offer a distinct advantage in handling the complex, non-linear relationships and high-dimensional data typical of diabetes pathology.

Consequently, the implementation of these sophisticated analytical tools does not merely represent a technical upgrade but a fundamental transformation in patient care strategy. By enabling more accurate risk assessment and treatment optimization, these models facilitate a move from reactive treatment to proactive, personalized medicine. Future research should focus on refining hybrid models that combine the interpretability of statistical inference with the computational power of deep learning, thereby bridging the gap between theoretical mathematics and clinical application. Ultimately, the integration of these advanced forecasting methods is essential for navigating the complexities of diabetes treatment and improving long-term patient outcomes.

Acknowledgments

The author would like to express his sincere gratitude to the reviewers for their valuable comments and technical advic.

Funding

[DOI: 10.18502/ijdo.v18i2.21648]

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

The author confirms sole responsibility for the conceptualization, methodology, data analysis, and the writing and editing of the manuscript.

[DOI: 10.18502/ijdo.v18i2.21648]

References

1. Ley C, Martin RK, Pareek A, Groll A, Seil R, Tischer T. Machine learning and conventional statistics: making sense of the differences. Knee Surgery, Sports Traumatology, Arthroscopy. 2022;30(3):753-7.
2. Chen Z, Huang C, Zhou Z, Zhang Y, Xu M, Tang Y, et al. A nonlinear associations of metabolic score for insulin resistance index with incident diabetes: A retrospective Chinese cohort study. Frontiers in Clinical Diabetes and Healthcare. 2023;3:1101276.
3. Alanazi BS. A comparative study of traditional statistical methods and machine learning techniques for improved predictive models. International Journal of Analysis and Applications. 2025;23:18.
4. Pakgohar A, Saffarzadeh M, Khalili M. The survey role of humanistic factor in incidence and intensity of road accident based on Logistic Regression and CART. Tehran: Applied Research Office of Traffic Police. 2008;13(5):49-66.
5. Pakgohar A, Khalili M, Safarzadeh M. Road traffic accident reduction via GLM, CRT, LR regression models. 2010;12(146):77-106.(in Persian)
6. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters. 2023;10(1-2):1-0.
7. GR A, Mary X A, George S T, Sagayam K M, Fernandez-Gamiz U, Günerhan H, et al. Analysis of diabetes disease using machine learning techniques: A review. Journal of Information Technology Management. 2023;15(4):139-59.
8. Wooldridge JM. Econometric analysis of cross section and panel data MIT press. Cambridge, ma. 2002;108(2):245-54.
9. Obster F, Ciolacu MI, Humpe A. Balancing Predictive Performance and Interpretability in Machine Learning: A Scoring System and an Empirical Study in Traffic Prediction. IEEE Access. 2024;12:195613-28.
10. Barbierato E, Gatti A. The challenges of machine learning: A critical review. Electronics. 2024;13(2):416.
11. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010;33:1-22.
12. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: springer; 2001.
13. Pakgohar A, Khalili M. Investigation of sample size in qualitative sampling methods. Popularization of Science. 2021;12(1):270-97.
14. Pakgohar A. The Role of Sample Size on Interpretation of the Result in Applied Research A Study on the Analysis of Regression Models. Methodology of Social Sciences and Humanities. 2023;29(114):19-34.
15. Pakgohar A. Sample Size calculation based on research Approaches in Animal Studies. Journal of Biostatistics and Epidemiology. 2023;9(4):474-83.
16. Pakgohar A, Mehrannia H. Sample size calculation in clinical trial and animal studies. Iranian Journal of diabetes and Obesity.2024,16(1): 42-50.
17. Pakgohar, A, and Fazli M.A. (2024) Artificial Intelligence, Statistics and Big Data in Medicine and Healthcare. 5th International Conference on Software Computing.(in Persian). https://civilica.com/doc/1966976/
18. Akinrinmade AO, Adebile TM, Ezuma-Ebong C, Bolaji K, Ajufo A, Adigun AO, et al. Artificial intelligence in healthcare: perception and reality. Cureus. 2023;15(9):e45594.
19. Guan Z, Li H, Liu R, Cai C, Liu Y, Li J, et al. Artificial intelligence in diabetes management: advancements, opportunities, and challenges. Cell Reports Medicine. 2023;4(10):101213.
20. Smith, J. K. Advanced statistical methods in healthcare research. Oxford University Press. 2020. https://link.springer.com/book/10.1007/978-981-15-8210-3
21. Zabbah I, Eskandari A, Sardari Z, Noghandi A. Diagnosis of diabetes using artificial neural network and neuro-fuzzy approach. Journal of Health and Biomedical Informatics .2018;5(2):274-85.(in Persian)
22. Bukhari MM, Alkhamees BF, Hussain S, Gumaei A, Assiri A, Ullah SS. An improved artificial neural network model for effective diabetes prediction. Complexity. 2021;2021(1):5525271.
23. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x
24. Johnson, A. E. W., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35
25. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. The New England Journal of Medicine, 378(11), 981-983. https://doi.org/10.1056/NEJMp1714229
26. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 2019;25(1):44-56.

[DOI: 10.18502/ijdo.v18i2.21648]

27. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: springer; 2013. https://doi.org/10.1007/978-1-4614-7138-7
28. Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. Global sensitivity analysis: the primer. John Wiley & Sons; 2008.https://doi.org/10.1002/9780470725184
29. Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 2019. https://doi.org/10.1002/9781119482260
30. Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall/CRC; 1994. https://doi.org/10.1201/9780429246593
31. Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. New York: springer; 2006. https://doi.org/10.1007/978-0-387-31073-2
32. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, et al. Apache spark: a unified engine for big data processing. Communications of the ACM. 2016;59(11):56-65. https://doi.org/10.1145/2934664
33. Maaten LV, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9:2579-605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[DOI: 10.18502/ijdo.v18i2.21648]