JESPAR

Journal of Exercise Sciences
& Physical Activity Reviews

ORIGINAL ARTICLE

Predictive model for glycemic control in patients with diabetes mellitus

Fatma Hilal Yagin[1*] | **Georgian Badicu**[2]

[1]Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, Türkiye
[2]Department of Physical Education and Special Motricity, Transilvania University of Brasov, 500068 Brasov, Romania

ABSTRACT

Diabetes mellitus, a chronic disease characterized by high blood sugar levels, necessitates effective glycemic control to prevent severe complications such as damage to the heart, blood vessels, eyes, kidneys, and nerves. This study aims to utilize machine learning techniques to predict glycemic control among a open Access dataset of 77,723 newly diagnosed diabetic patients in Istanbul. By employing a logistic regression model, the study identifies key features influencing glycemic control, enhancing model interpretability for clinicians. The model demonstrates robust performance with an accuracy of 0.825, precision scores of 0.86 (positive class) and 0.76 (negative class), recall values of 0.86 (positive class) and 0.77 (negative class), and corresponding F1 scores. Feature importance analysis reveals HbA1c as the dominant predictor, significantly surpassing other variables. These findings provide critical insights into the application of machine learning in diabetes management, highlighting the pivotal role of HbA1c in glycemic control prediction.

Keywords: Diabetes mellitus, glycemic control, machine learning, predictive modeling

***Corresponding:** Fatma Hilal Yagin; hilal.yagin@inonu.edu.tr
Journal home page: www.e-jespar.com
Academic Editor: Dr. Mehmet Gülü
https://doi.org/10.5281/zenodo.12601066

ARTICLE HISTORY
Received: 11 May 2024
Accepted: 27 May 2024
Published: 01 July 2024

INTRODUCTION

Diabetes mellitus, commonly known as diabetes, is a chronic disease characterized by high levels of blood sugar (or blood glucose) that, over time, causes serious damage to the heart, blood vessels, eyes, kidneys, and nerves. Effectively managing blood sugar levels is crucial to preventing these complications (Kaul, Tarr, Ahmad, Kohner, & Chibber, 2013).

Glycemic control, the typical measure of how well blood sugar levels are managed, is therefore a central focus in diabetes care (B. P. Kovatchev, 2017). Glycemic control is influenced by a multitude of factors, from demographics and lifestyle choices to clinical and pharmacological interventions. Understanding these factors and their relative importance can significantly improve management strategies for people with diabetes (Cheng, Wang, Lim, & Wu, 2019; B. Kovatchev, 2019).

Machine learning models provide a powerful way to analyze large data sets to uncover patterns and relationships that may not be immediately obvious with traditional statistical methods. In recent years, the application of machine learning techniques in healthcare has increased, providing valuable insights into disease prediction, patient outcomes, and treatment effectiveness. These advanced analytical methods enable the processing of complex, high-dimensional data, enabling the identification of the most critical variables affecting glycemic control among diabetic patients (Dagliati et al., 2018; Greener, Kandathil, Moffat, & Jones, 2022; L'heureux, Grolinger, Elyamany, & Capretz, 2017; Lai, Huang, Keshavjee, Guergachi, & Gao, 2019; Najafabadi et al., 2015).

This study aims to utilize machine learning approaches to predict glycemic control in a dataset of 77,723 patients and identify key features that affect this condition. By applying a logistic regression model, we aim to determine the most effective model to predict glycemic control and analyze the importance of different features. Variable importance is important to increase the interpretability of the model, thus translating the complex outputs of the machine learning model into actionable insights for clinicians and healthcare professionals. Our findings provide important insights into the application of machine learning in diabetes and glycemic control management by providing a detailed overview of how different factors contribute to glycemic control.

METHODS

Participant and Data

An open access diabetes dataset was used in this investigation. The newly diagnosed diabetic patients in Istanbul in 2017 were included in the data used to evaluate glycemic control three years following diagnosis. Patients were classified into two groups based on their HbA1c level profiles: under control (last two HbA1c values below 7) and poorly controlled. A total of 105 variables were taken out and utilized as independent variables for 77,723 patients from the e-Nabız system (Mendeley dataset).

Data Preprocessing

In order to prepare the dataset for analysis and modeling, missing value imputation, training/test set separation and normalization processes were applied. Missing values were imputed with the mean of the relevant column. Data were divided into 70% training and 30% test set. Variables were normalized using standard scaler (García, Luengo, & Herrera, 2015).

Model Training, Evaluation and Variable Importance

Logistic regression classification algorithm (Dreiseitl & Ohno-Machado, 2002) was used in the modeling phase for glycemic control prediction. Data was trained using the training set and model performance was evaluated on the test set. The model's accuracy and classification report (precision, recall, f1-score) were calculated to evaluate the prediction performance (Yacouby & Axman, 2020). Then, in order to examine the contribution of variables to glycemic control prediction, the most important 20 variables were visualized using the variable importance graph (Genuer, Poggi, & Tuleau-Malot, 2010).

RESULTS

Confusion matrix for logistic regression model is presented in Table 1 and performance metrics are presented in Table 2. The model exhibits strong performance across various key metrics, with an accuracy of 0.825, indicating it correctly predicts outcomes 82.5% of the time, reflecting high overall reliability. Precision scores of 0.86 for the positive class and 0.76 for the negative class, with a mean of 0.81, demonstrate that the model accurately identifies 86% of positive predictions and 76% of negative predictions. Similarly, recall values of 0.86 for the positive class and 0.77 for the negative class, averaging to 0.81, show the model's capability to correctly identify 86% of actual positive cases and 77% of actual negative cases. The F1 scores, also 0.86 and 0.77 for the positive and negative classes respectively, with a mean of 0.81, further confirm the balanced performance between precision and recall. The confusion matrix, with 12,567 true positives, 6,754 true negatives, 1,975 false positives, and 2,022 false negatives, provides a detailed breakdown of the model's predictions, underscoring its effectiveness in classification tasks (Table 1, and Table 2).

	Predicted Negative	Predicted Positive
Actual Negative	6754	1975
Actual Positive	2022	12567

Table 1. Confusion matrix for logistic regression model for glycemic control prediction

Metric	Description	Value

Accuracy	Overall correctness of the model; the proportion of true results (both true positives and true negatives) among the total number of cases examined.	0.825
Precision	The proportion of true positive results in all positive predictions made by the model. Precision indicates how many of the positive predictions made by the model were actually correct.	0.86 (Positive class) / 0.76 (Negative class), mean 81
Recall (Sensitivity)	The proportion of true positive results in all actual positives. Recall indicates how many of the actual positive cases the model was able to identify correctly.	0.86 (Positive class) / 0.77 (Negative class), mean 81
F1 Score	The harmonic mean of precision and recall, providing a single metric that balances both concerns.	0.86 (Positive class) / 0.77 (Negative class), mean 81
Confusion Matrix	A table used to describe the performance of the classification model by showing the actual vs. predicted classifications.	TP: 12567, TN: 6754, FP: 1975, FN: 2022

Table 2. Performance of the logistic regression model for glycemic control prediction on the test set

The feature importance analysis reveals that HbA1c is the most significant predictor, with a relative importance close to 0.9, far surpassing all other features. The second most important feature is HbA1c_change, though its importance is substantially lower, indicating that changes in HbA1c levels also play a crucial role but to a lesser extent. Other features like gliklazid, insulin glargin, insulin aspart, and age show minimal importance, suggesting they have a relatively minor impact on the model's predictions. Features such as insulin lispro, metformin hcl, HDL, and insulin detemir, along with others like obesity, creatinine, sitagliptin, glimepirid, dapagliflozin, repaglinide, LDL, cholesterol, sex, and thyroid disease, exhibit negligible relative importance, indicating that they contribute very little to the predictive power of the model compared to HbA1c. This highlights the dominant role of HbA1c in influencing the model's outcomes, underscoring its critical importance in the context of the analysis.

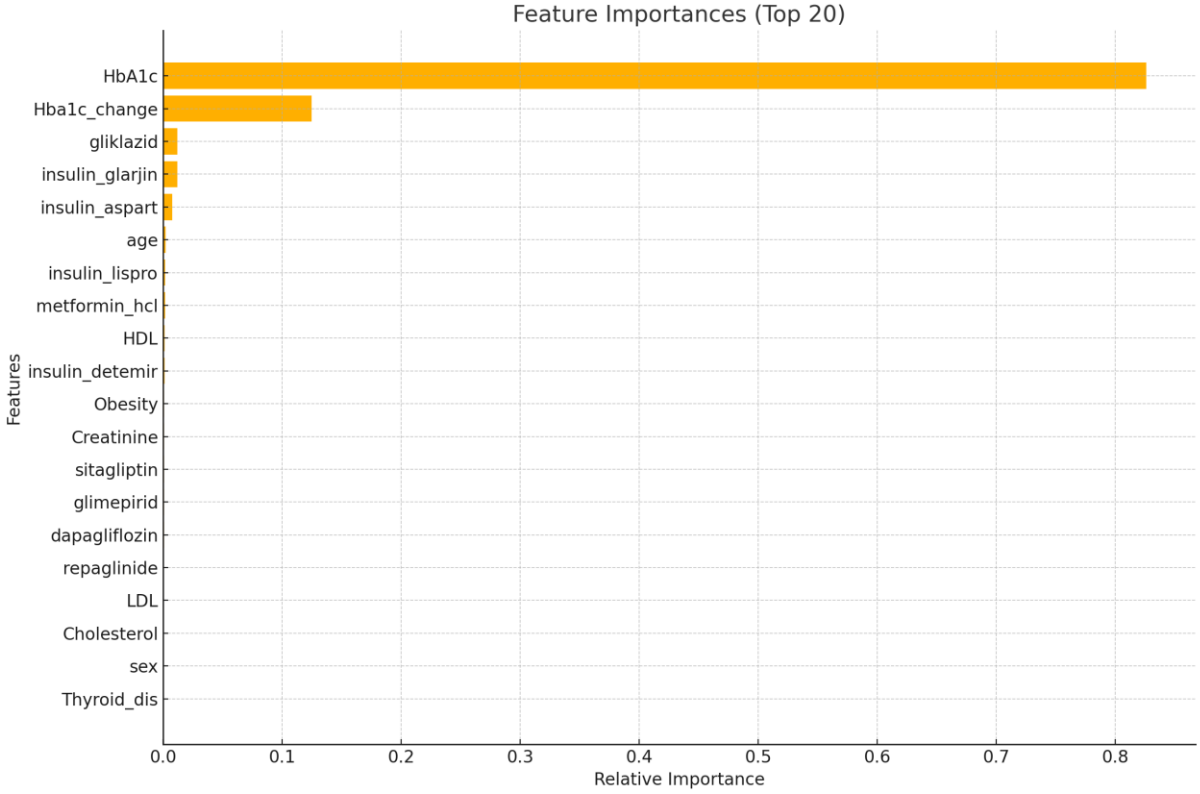

Figure 1. Importance plot of the 20 most important variables for glycemic control prediction.

DISCUSSION

The results of this study underscore the efficacy of machine learning models, particularly logistic regression, in predicting glycemic control among diabetic patients. With an accuracy of 0.825, the model reliably predicts outcomes, affirming its potential utility in clinical settings. The high precision and recall scores for the positive class (0.86 each) indicate that the model is adept at correctly identifying patients with controlled blood sugar levels. Conversely, the slightly lower precision and recall for the negative class (0.76 and 0.77, respectively) suggest areas for improvement in identifying poorly controlled cases. The confusion matrix provides a detailed breakdown, with 12,567 true positives and 6,754 true negatives, alongside 1,975 false positives and 2,022 false negatives. This detailed performance evaluation highlights the model's strengths and areas needing refinement. The F1 scores, which balance precision and recall, further validate the model's balanced performance across both positive and negative classes.Feature importance analysis reveals that HbA1c is the most critical predictor, with a relative importance approaching 0.9, emphasizing its crucial role in managing diabetes. The second most important feature, HbA1c_change, though significantly less impactful, still plays a vital role in predicting glycemic control. Other features such as gliklazid, insulin glargin, and insulin aspart, alongside demographic variables like age, show minimal importance. This suggests that while these factors contribute to the model's predictions, their impact is considerably lower compared to HbA1c. Less significant features include various medications (e.g., insulin lispro, metformin hcl), lifestyle factors (e.g., obesity), and other biomarkers (e.g., HDL, creatinine), which exhibit negligible relative importance. These findings indicate that while these variables are part of the overall prediction model, their influence is marginal when compared to HbA1c levels. Overall, this study highlights the dominant role of HbA1c in predicting glycemic control, underscoring its importance in clinical practice. The insights gained from this model can inform better management strategies for diabetes, aiding healthcare providers in identifying critical factors that influence patient outcomes.

CONCLUSIONS

This study employed logistic regression to predict glycemic control in newly diagnosed diabetic patients. The model achieved strong performance with an accuracy of 0.825, demonstrating reliable predictions. Key metrics such as precision (0.86 for positive class, 0.76 for negative class), recall (0.86 for positive class, 0.77 for negative class), and F1 score (0.86 for positive class, 0.77 for negative class) underscore its balanced performance. Feature importance analysis highlighted HbA1c as the most influential predictor, emphasizing its critical role in determining glycemic outcomes. These findings underscore the significance of HbA1c monitoring in diabetes management.

Author Contributions

Conceptualization, F.H.Y. methodology, F.H.Y, G.B.; formal analysis, G.B.; investigation, G.B.; data curation, F.Y.H.; writing–original draft preparation, F.H.Y, G.B.; writing–review and editing, F.H.Y, G.B.

Informed Consent Statement:

The research was conducted in line with the Declaration of Helsinki.

Acknowledgments:

We would like to thank all participants who took part in the research.

Funding:

This research was not funded by any institution or organization.

Conflicts of Interest:

The authors declare that no conflicts interest.

REFERENCES

Cheng, L. J., Wang, W., Lim, S. T., & Wu, V. X. (2019). Factors associated with glycaemic control in patients with diabetes mellitus: a systematic literature review. *Journal of clinical nursing, 28*(9-10), 1433-1450.

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., . . . Bellazzi, R. (2018). Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology, 12*(2), 295-302.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics, 35*(5-6), 352-359.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72): Springer.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters, 31*(14), 2225-2236.

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature reviews Molecular cell biology, 23*(1), 40-55.

Kaul, K., Tarr, J. M., Ahmad, S. I., Kohner, E. M., & Chibber, R. (2013). Introduction to diabetes mellitus. *Diabetes: an old disease, a new insight*, 1-11.

Kovatchev, B. (2019). Glycemic variability: risk factors, assessment, and control. *Journal of diabetes science and technology, 13*(4), 627-635.

Kovatchev, B. P. (2017). Metrics for glycaemic control—from HbA1c to continuous glucose monitoring. *Nature Reviews Endocrinology, 13*(7), 425-436.

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access, 5*, 7776-7797.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC endocrine disorders, 19*, 1-9.

Mendeley. (n.d.). Dataset: Chronic fatigue syndrome metabolomics. Mendeley Data. Retrieved June 30, 2024, from https://data.mendeley.com/datasets/rr4rzzrjfc/2

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of big Data, 2*, 1-21.

Yacouby, R., & Axman, D. (2020). *Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models.* Paper presented at the Proceedings of the first workshop on evaluation and comparison of NLP systems.