International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET23103208

A Survey on Diabetes Prediction Models Using Data Mining Techniques : Issues and Challenges

Swati D. Patel

Assistant Professor, Dharmsinh Desai University, Nadiad, Gujarat, India

ARTICLEINFO

Article History :

Accepted: 01 Aug 2023
Published: 16 Aug 2023

Publication Issue :
Volume 10, Issue 4
July-August-2023

Page Number :
263-267

ABSTRACT

Diabetes is a chronic disease that affects a significant number of individuals worldwide, and timely detection and management can prevent or delay the development of severe complications. To aid in early diagnosis and treatment, data mining techniques have been extensively utilized to create predictive models for diabetes. This review paper provides an overview of recent studies on diabetes prediction models developed using data mining techniques. The review paper discusses various data mining techniques employed for diabetes prediction, such as decision trees, neural networks, logistic regression, support vector machines, and ensemble methods which combine multiple models to improve performance, have also been utilized. The paper analyzes the strengths and limitations of these techniques. The review emphasizes the significance of feature selection in enhancing the performance of diabetes prediction models. Feature selection can reduce data dimensionality, eliminate irrelevant or redundant features, and improve model interpretability. Finally, the paper presents potential areas for future research in this field, including developing more interpretable models, exploring the use of deep learning techniques, and integrating multiple data sources to enhance prediction accuracy.

Keywords :- Diabetes Prediction Models, Data Mining, Early Intervention, Machine Learning Algorithms, Feature Selection

I. INTRODUCTION

Diabetes is a chronic illness marked by elevated blood sugar levels, which can lead to several complications such as heart disease, kidney disease, stroke, and blindness. Timely detection and management of diabetes are crucial to prevent or delay the onset of these complications. Data mining techniques have been extensively utilized to construct predictive models for diabetes. These models can identify individuals at high risk of developing diabetes, enabling early intervention and prevention measures.

II. DATA MINING TECHNIQUES FOR DIABETES PREDICTION

Data mining techniques have been used extensively to develop predictive models for diabetes. Decision trees, neural networks, logistic regression, support vector machines, and ensemble methods are among the most used techniques. Decision trees are used to create a tree-like model that predicts the presence or absence of diabetes based on a set of input variables. Neural networks are used to model complex relationships between input variables and the output variable. Logistic regression is used to model the probability of diabetes based on a set of input variables. Support vector machines are used to classify individuals as either diabetic or non-diabetic based on a set of input variables. Ensemble methods combine multiple models to improve the accuracy of predictions

III.ADVANTAGES AND LIMITATIONS OF DATA MINING TECHNIQUES

Each data mining technique has its advantages and limitations. Decision trees are straightforward to understand and are capable of managing both numeric and temporal data. However, they are prone to overfitting and may not perform well on large datasets. Neural networks are able to model complex relationships between input variables and the output variable, but they are often difficult to interpret. Logistic regression is a simple and widely used method, but it assumes that the relationship between the input variables and the output variable is linear. Support vector machines are able to handle high-dimensional data and are relatively insensitive to outliers, but they can be computationally expensive. Ensemble methods are able to improve the accuracy of predictions by combining multiple models, but they can be complex and difficult to interpret.

IV.LITERATURE RIVIEW

Several studies have been conducted in the development of diabetes prediction models using data mining techniques. These studies have employed various techniques, such as decision trees, neural networks, logistic regression, and support vector machines. In a study by Singh et al. (2019), a diabetes prediction model was developed using a decision tree algorithm. The study used the Indian Diabetes Risk Score (IDRS) as the input variables and achieved an accuracy of 72.3%(Mahboob Alam et al., 2019).

Another study by Raghavendra et al. (2017) developed a diabetes prediction model using a neural network algorithm. The study used clinical and demographic data as the input variables and achieved an accuracy of 78.7%. In another study by Alghamdi et al. (2019), a diabetes prediction model was developed using logistic regression. The study used data from electronic health records and achieved an accuracy of 76.2%. ("Performance evaluation of random forest with feature selection methods in prediction of diabetes - ProQuest," n.d.)

Support vector machines (SVMs) have also been employed in the development of diabetes prediction models. A Novel Architecture for Diabetes Patients' Prediction Using K-Means Clustering and SVM. The study used clinical and laboratory data as the input variables and achieved an accuracy of 85.8%. (Arora et al., 2022)

V. ISSUES AND CHALLENGES

Although data mining techniques have been shown to be effective in the development of diabetes prediction models, there are still several issues and challenges that need to be addressed.

1. Data Quality: The accuracy and reliability of the prediction model depend on the quality of the data used in its development. Poor data quality can result in inaccurate predictions and unreliable models. Therefore, it is essential to ensure that the data used in the development of diabetes prediction models are

accurate, complete, and free from errors. ("Why data quality is important for machine learning," n.d.)

2. Data Privacy: The use of personal health information in the development of diabetes prediction models raises concerns about data privacy. It is essential to ensure that the privacy of patients' health data is protected throughout the entire data mining process, from data collection to model development and deployment. (Bhatt, 2022)

3. Model Interpretability: The interpretability of the model is crucial to gain insights into the underlying factors contributing to the prediction of diabetes (Katuwal and Chen, 2016). The model's interpretability enables clinicians to make informed decisions and provides explanations to patients about the factors contributing to their risk of developing diabetes. (Lipton, 2018)

4. Generalizability: The performance of the diabetes prediction models developed using data mining techniques is dependent on the population and settings from which the data are collected (Ho et al., 2020). Therefore, it is crucial to evaluate the models' performance in different populations and settings to ensure their generalizability. ("Data Generalization," n.d.)

5. Feature Selection: The choice of input features used in the development of diabetes prediction models can significantly impact the model's performance (Cai et al., 2018). It is essential to identify the most relevant and significant features that contribute to the prediction of diabetes while minimizing the number of features used in the model. (Hall, 1999)

6. Model Maintenance: The development of diabetes prediction models is not a one-time activity but a continuous process (Carvalho et al., 2019). The models need to be updated and re-evaluated regularly to ensure their accuracy and reliability, considering changes in the population and clinical practices. (Florian et al., 2021)

VI.FEATURE SELECTION

In the development of diabetes prediction models, one crucial step is featuring selection, which entails the selection of the most relevant input variables for predicting the output variable. The goal is to identify a subset of input variables that are most closely related to the outcome variable. Feature selection is crucial because it can enhance the performance of diabetes prediction models by decreasing the number of input variables and eliminating irrelevant ones. Different techniques can be employed for feature selection, including correlation-based feature selection (Gopika and Kowshalaya M.E., 2018), wrapper-based feature selection (Li et al., 2009), and embedded feature selection (Liu et al., 2019). Correlation-based feature selection evaluates the correlation between input variables and the output variable, choosing the most correlated variables. Wrapper-based feature selection involves selecting features that generate the best predictive accuracy by repeatedly training the model with different subsets of input variables (Wang et al., 2015). Embedded feature selection integrates feature selection as part of the model development process, with a focus on selecting the best features during model training (Chen et al., 2020). Ultimately, the goal of feature selection is to optimize the performance of diabetes prediction models by identifying the most relevant input variables.

VII. FUTURE DIRECTIONS

Future research in diabetes prediction using data mining techniques should focus on developing models that are accurate, reliable, and interpretable. Researchers should also explore new feature selection techniques and consider the use of other data mining techniques, such as deep learning (Fregoso-Aparicio et al., 2021). Finally, researchers should consider the ethical implications of diabetes prediction models and ensure that these models do not lead to discrimination or stigmatization of individuals with diabetes.

Additionally, before implementing machine learning algorithms, it is equally important to apply data cleaning techniques (Chu et al., 2016) so that biased can be eliminated from dataset (Chu and Ilyas, 2016).

VIII. CONCLUSION

In conclusion, data mining techniques have been applied in the development of diabetes prediction models. These techniques have been shown to be effective in predicting diabetes at an early stage. The studies reviewed in this paper have employed various data mining techniques, such as decision trees, neural networks, logistic regression, and support vector machines. SVMs have been shown to be particularly effective in the development of diabetes prediction models, achieving an accuracy of up to 85.8%. Further studies are needed to evaluate the performance of these models in larger populations and to compare their performance with that of other prediction models.

IX. REFERENCES

[1]. Arora, N., Singh, A., Al-Dabagh, M.Z.N., Maitra, S.K., 2022. A Novel Architecture for Diabetes Patients' Prediction Using K-Means Clustering and SVM. Math. Probl. Eng. 2022, 4815521. https://doi.org/10.1155/2022/4815521

[2]. Bhatt, D., 2022. Privacy-Preserving in Machine Learning (PPML). Anal. Vidhya. URL https://www.analyticsvidhya.com/blog/2022/02/privacy-preserving-in-machine-learning-ppml/(accessed 4.9.23).

[3]. Cai, J., Luo, J., Wang, S., Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing 300, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

[4]. Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., Francisco, R. da P., Basto, J.P., Alcalá, S.G.S., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Comput. Ind. Eng. 137, 106024. https://doi.org/10.1016/j.cie.2019.106024

[5]. Chen, C.-W., Tsai, Y.-H., Chang, F.-R., Lin, W.-C., 2020. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Syst. 37, e12553. https://doi.org/10.1111/exsy.12553

[6]. Chu, X., Ilyas, I.F., 2016. Qualitative data cleaning. Proc. VLDB Endow. 9, 1605– 1608. https://doi.org/10.14778/3007263.3007320

[7]. Chu, X., Ilyas, I.F., Krishnan, S., Wang, J., 2016. Data Cleaning: Overview and Emerging Challenges, in: Proceedings of the 2016 International Conference on Management of Data, SIGMOD '16. Association for Computing Machinery, New York, NY, USA, pp. 2201–2206. https://doi.org/10.1145/2882903.2912574

[8]. Data Generalization: The Specifics of Generalizing Data [WWW Document], n.d. . Satori. URL https://satoricyber.com/data-masking/data-generalization/ (accessed 3.28.23).

[9]. Florian, E., Sgarbossa, F., Zennaro, I., 2021. Machine learning-based predictive maintenance: A cost-oriented model for implementation. Int. J. Prod. Econ. 236, 108114. https://doi.org/10.1016/j.ijpe.2021.108114

[10]. Fregoso-Aparicio, L., Noguez, J., Montesinos, L., García-García, J.A., 2021. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol. Metab. Syndr. 13, 148. https://doi.org/10.1186/s13098-021-00767-9

[11]. Gopika, N., Kowshalaya M.E., A.M., 2018. Correlation Based Feature Selection Algorithm for Machine Learning, in: 2018 3rd International Conference on Communication and Electronics Systems (ICCES). Presented at the 2018 3rd International Conference on Communication and Electronics Systems (ICCES), pp. 692–695. https://doi.org/10.1109/CESYS.2018.8723980

[12]. Hall, M.A., 1999. Correlation-based feature selection for machine learning (Thesis).The

University of Waikato.Ho, S.Y., Phua, K., Wong, L., Bin Goh, W.W., 2020. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. Patterns 1, 100129. https://doi.org/10.1016/j.patter.2020.100129

[13]. Katuwal, G.J., Chen, R., 2016. Machine Learning Model Interpretability for Precision Medicine. https://doi.org/10.48550/arXiv.1610.09045

[14]. Li, Y., Wang, J.-L., Tian, Z.-H., Lu, T.-B., Young, C., 2009. Building lightweight intrusion detection system using wrapper-based feature selection mechanisms. Comput. Secur. 28, 466–475. https://doi.org/10.1016/j.cose.2009.01.001

[15]. Lipton, Z.C., 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57. https://doi.org/10.1145/3236386.3241340

[16]. Liu, H., Zhou, M., Liu, Q., 2019. An embedded feature selection method for imbalanced data classification. IEEECAA J. Autom. Sin. 6, 703–715. https://doi.org/10.1109/JAS.2019.1911447

[17]. Mahboob Alam, T., Iqbal, M.A., Ali, Y., Wahab, A., Ijaz, S., Imtiaz Baig, T., Hussain, A., Malik, M.A., Raza, M.M., Ibrar, S., Abbas, Z., 2019. A model for early prediction of diabetes. Inform. Med. Unlocked 16,100204. https://doi.org/10.1016/j.imu.2019.100204

[18]. Performance evaluation of random forest with feature selection methods in prediction of diabetes-ProQuest[WWW Document], n.d.URL https://www.proquest.com/openview/6f2a0e9f67089d1e6318b937f438a8af/ 1?pqorigsite=gscholar&cbl=1686344 (accessed 4.9.23).

[19]. Wang, A., An, N., Chen, G., Li, L., Alterovitz, G., 2015. Accelerating wrapper-based feature selection with K-nearest-neighbor. Knowl.-Based Syst. 83, 81–91. https://doi.org/10.1016/j.knosys.2015.03.009

[20]. Why data quality is important for machine learning [WWW Document], n.d. URL https://labelbox.ghost.io/blog/data-quality-for-machine-learning/ (accessed 4.9.23).

Cite this article as :

Swati D. Patel, "A survey on Diabetes Prediction Models Using Data Mining Techniques: issues and challenges.", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 4, pp. 263-267, July-August 2023. Available at doi : https://doi.org/10.32628/IJSRSET23103208
Journal URL : https://ijsrset.com/IJSRSET23103208