International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I7P108

Volume 72 Issue 7, 69-73, July 2024
© 2024 Seventh Sense Research Group®

Original Article

Improving Lifestyle Choices with Diabetes Prediction with Help of R Programming

Saba Sultana[1], Vishal Patel[2]

[1]*Senior Software Engineer, PA, USA.*
[2]*Technical Architect, NJ, USA.*

[1]*Corresponding Author : ms.sabasultana@gmail.com*

Received: 14 May 2024 Revised: 25 June 2024 Accepted: 15 July 2024 Published: 30 July 2024

Abstract - *Diabetes prevalence is rising globally, demanding innovative preventative measures. This study explores the use of R programming for diabetes prediction to empower individuals with personalized lifestyle modifications. We develop a predictive model in R using a relevant dataset, identifying key risk factors through analysis. Based on the predicted risk, we propose targeted lifestyle changes for individuals, promoting preventative healthcare. This data-driven approach fosters informed decision-making, empowering individuals to adopt healthier habits and potentially mitigating diabetes development. This research contributes to the field of preventative healthcare by demonstrating the potential of R programming in tailoring lifestyle choices based on predicted diabetes risk.*

Keywords - *Diabetes, R programming, Lifestyle Modification, BMI, IDF.*

1. Introduction

Diabetes mellitus, a chronic metabolic disorder characterized by elevated blood sugar levels, has become a global health crisis. According to the International Diabetes Federation (IDF), in 2021, an estimated 537 million adults (aged 20-79 years) were living with diabetes, and this number is projected to rise to 783 million by 2045 [1]. This alarming increase translates to a significant burden on individuals and healthcare systems worldwide. Diabetes can lead to severe complications like heart disease, stroke, blindness, and kidney failure, significantly impacting quality of life and life expectancy [2]. Early detection of diabetes is crucial for effective management and potentially preventing its debilitating complications. Early intervention allows for timely treatment adjustments lifestyle modifications, and potentially even delaying or preventing the onset of the disease [3]. Lifestyle choices, particularly a healthy diet, regular physical activity, and maintaining a healthy weight, have been demonstrated to significantly reduce the risk of developing type 2 diabetes, the most common form of the disease [4]. Recent advancements in data science and machine learning offer promising tools for diabetes prevention. Predictive models, built using historical data on individuals with and without diabetes, can analyze various factors and identify individuals at a higher risk of developing the disease [5]. This risk assessment allows for early intervention and the implementation of personalized preventative strategies. R is a powerful open-source programming language gaining significant traction in the field of healthcare research [6]. Its extensive suite of statistical libraries and data visualization tools makes it ideal for building and analyzing predictive models [7]. The open-source nature of R allows for transparency, collaboration, and wide adoption within the research community. This paper explores the potential of R programming for building a diabetes prediction model. By leveraging R's capabilities to analyze relevant diabetes data a predictive model is developed which identifies key risk factors associated with the disease. Based on the predicted risk, personalized lifestyle modifications can be proposed so that individuals can adopt them to prevent or delay the onset of diabetes potentially.

2. Diabetes: A Global Health Crisis

Diabetes, a chronic metabolic disorder characterized by elevated blood sugar levels, has emerged as a significant global health issue, affecting millions worldwide. According to the World Health Organization (WHO), the prevalence of diabetes has been steadily rising, with an estimated 422 million adults living with the condition in 2014, and projections suggest this number will reach 642 million by 2040 [8]. This epidemic status underscores the critical need for effective preventive measures and management strategies. Predictive modeling has emerged as a promising approach for assessing the risk of developing diabetes. By analyzing various factors such as demographic information, medical history, and lifestyle habits, predictive models can estimate an individual's likelihood of developing diabetes within a specified timeframe. These models not only enable early identification of at-risk individuals but also facilitate targeted interventions and resource allocation for prevention and management efforts [9]. The development and deployment of predictive models for diabetes risk assessment hold immense potential for public health initiatives. By leveraging advanced analytics and machine learning techniques, healthcare providers and policymakers can identify high-risk individuals, tailor interventions to their specific needs, and allocate resources more efficiently.

Moreover, predictive models can contribute to the development of personalized healthcare strategies, empowering individuals to make informed lifestyle choices and adopt preventive measures to reduce their risk of developing diabetes. Overall, predictive modeling represents a valuable tool in the global fight against diabetes, offering the potential to improve health outcomes, reduce healthcare costs, and alleviate the burden of this chronic disease on individuals and society as a whole.

3. Literature Review

Lifestyle choices play a pivotal role in the management and prevention of diabetes. Factors such as diet, physical activity levels, and smoking habits significantly influence an individual's risk of developing the condition. Adopting a healthy lifestyle, characterized by a balanced diet, regular exercise, and avoidance of tobacco products, can help control blood sugar levels, prevent complications, and reduce the overall burden of diabetes [10].

Recent literature has highlighted the potential of precision diabetes medicine, emphasizing the optimization of therapy through patient-level biomarker data to achieve more effective, cost-efficient, and safer prevention, treatment, and potentially even cure for diabetes. While precision medicine often focuses on pharmacotherapy, there's growing recognition of its relevance to lifestyle interventions in diabetes management. Four main avenues have been proposed: firstly, predicting susceptibility to adverse lifestyle exposures; secondly, stratifying type 2 diabetes into subclasses to tailor specific lifestyle interventions; thirdly, discovering prognostic biomarkers to guide timing and intensity of lifestyle interventions; and finally, predicting treatment response [11].

Social media analytics present an untapped resource for understanding the lifestyle choices of individuals with diabetes, yet research in this area remains limited. It explored the potential of leveraging social media data, employing sentiment analysis and unsupervised topic modeling to uncover lifestyle-related discussions among individuals dealing with diabetes. Their findings underscore the need for predictive modeling approaches to quantify insights gleaned from social media data and improve population health outcomes [12]. In rural Indonesia, [13] investigated factors influencing healthy lifestyle behaviors among patients with type 2 diabetes. Using the extended health belief model along with demographic characteristics, clinical factors, and diabetes knowledge, they found that these factors collectively accounted for a significant proportion of variance in healthy lifestyle behaviors. The study highlights the importance of understanding socio-cultural factors and health beliefs in shaping lifestyle choices among diabetes patients in diverse settings.

Longitudinal studies, such as that conducted by [14] in Japan, have identified trajectory patterns of combined lifestyles over time and their association with diabetes risk. Their findings emphasize the importance of improving and maintaining health-related lifestyles to prevent diabetes, suggesting that sustained lifestyle modifications can significantly reduce diabetes risk. Machine learning techniques have been increasingly employed to predict diabetes risk based on demographic and lifestyle information. It developed a classification model to predict chronic diabetes using a person's demography and lifestyle attributes, achieving an accuracy of 80% [15].

As reviewed various machine learning algorithms for diabetes prediction, demonstrating their effectiveness in risk classification using datasets such as the Pima Indians Diabetes Database [15,16]. In China, applied decision trees, random forests, and neural network models were applied to predict diabetes mellitus using hospital examination data. They found that random forest achieved the highest accuracy, underscoring the potential of machine learning in diabetes prediction. These studies collectively highlight the growing interest in leveraging advanced analytics and machine learning to improve diabetes prediction and management strategies [17].

4. Materials and Methods

The dataset typically comprises variables such as gender, age, hypertension, heart disease, smoking history, Body Mass Index (BMI), HbA1c level, blood glucose level, and diabetes status. Challenges specific to this dataset may include missing values, outliers, and imbalanced classes, which necessitate thorough data cleaning and preprocessing steps. Data cleaning involves handling missing values through imputation or deletion, identifying and addressing outliers that may distort analysis results, and ensuring data consistency and integrity.

Feature selection is crucial for identifying relevant predictors and reducing dimensionality, which can enhance model performance and interpretability. Transformation techniques such as normalization or standardization may be applied to ensure that variables are on the same scale, facilitating model convergence and interpretation. R programming provides a comprehensive toolkit for data preprocessing, offering various packages and functions for data cleaning, feature selection, and transformation, thereby enabling researchers to prepare high-quality data for diabetes prediction models.

4.1. Diabetes Dataset

Here's a description of each column in the dataset:

Table 2. Dataset description

Column	Description
Gender	The gender of the individual (e.g., male or female)
Age	The age of the individual in years
Hypertension	Indicates whether the individual has hypertension (1 for yes, 0for no)
Heart Diseases	Indicates whether the individual has heart disease (1 for yes, 0 for no)
Smoking History	Smoking history of the individual (e.g., smoker, non-smoker)
BMI	The Centers for Disease Control and Prevention report that 32% of white and 53% of black women are obese. Women with a body mass index (BMI) of 30 kg/m2 have 28 times greater risk of developing diabetes than do women of normal weight. The risk of diabetes is 93 times greater if the BMI is 35 kg/m2. [23]
HbA1c_level	A level below 5.7% is considered normal.
	A level between 5.7% and 6.4% indicates prediabetes, suggesting a higher risk of developing diabetes.
	A level of 6.5% or higher on two separate tests are used to diagnose diabetes [24].
Blood Glucose level	People with diabetes verses No Diabetes (Boxplot their blood Glucose)
Diabetes	Indicates whether the individual has diabetes (1 for yes, 0 for no)

Table 2. Diabetes counts

Diabetes	Counts
0	91500
1	8500

5. Results and Discussion

5.1. Implementation

To implement diabetes prediction using R and the Random Forest algorithm, follow these steps. First, prepare your dataset by loading the data into R and performing any necessary preprocessing steps, such as handling missing values and scaling numeric features. Next, split the dataset into training and testing sets to evaluate the model's performance. Then, train the Random Forest model using the training data, specifying the number of trees and other hyperparameters. Once the model is trained, evaluate its performance on the testing data using metrics such as accuracy, precision, recall, and F1-score. Finally, interpret the results and assess the importance of features in predicting diabetes risk by analyzing the variable importance plot generated by the Random Forest model. This approach leverages the power of Random Forests in handling complex datasets and provides insights into the predictive factors contributing to diabetes risk, enabling informed decision-making for lifestyle modifications and healthcare interventions.

5.2. Results

The analysis involved implementing four machine learning algorithms using R programming – k-Nearest Neighbors (k-NN), Decision Tree, Random Forest, and Logistic Regression – to predict diabetes based on a dataset containing features such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level. Initially, the dataset was divided into training and testing sets to evaluate the performance of each algorithm. For each algorithm, the model was trained on the training set and then evaluated on the testing set using various performance metrics, including accuracy, precision, and recall. Accuracy represents the proportion of correctly predicted instances out of all instances, precision measures the proportion of true positive predictions out of all positive predictions, and recall indicates the proportion of true positive predictions out of all actual positive instances. The results, summarized in the table below, provide insights into the effectiveness of each algorithm in predicting diabetes, allowing for informed decision-making regarding the selection of the most suitable algorithm for diabetes prediction based on the dataset and performance metrics.

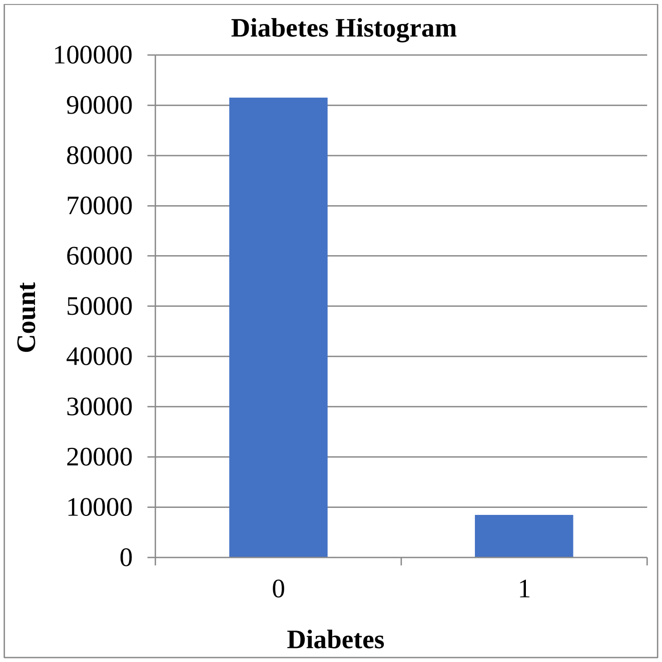

Fig. 1 Diabetes histogram

Table 3. Algorithm comparison

Algorithm	Accuracy	Precision	Recall
k-Nearest Neighbors	0.9493	0.8049	0.5363
Decision Tree	0.9464	0.6819	0.6979
Random Forest	0.9674	1.0	0.6183
Logistic Regression	0.9509	0.7993	0.5667

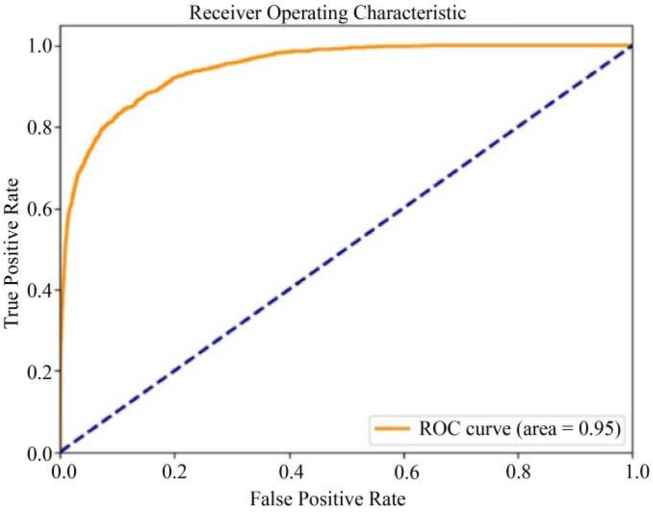

Fig. 2 Logistic regression

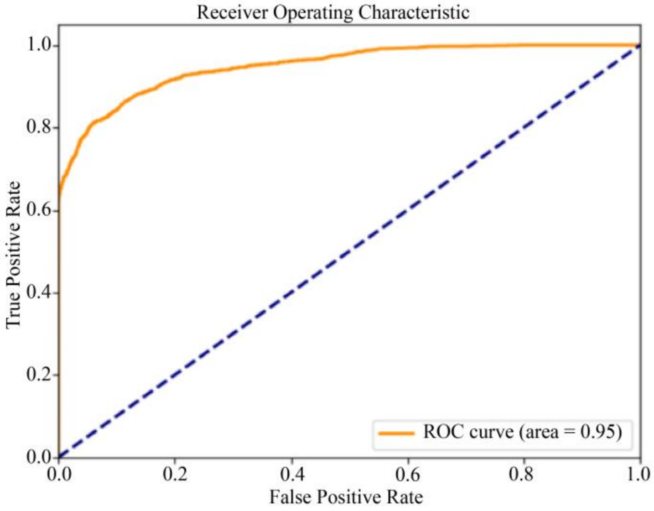

Fig. 3 Random forest classifier

In the evaluation of various machine learning algorithms for diabetes prediction, Random Forest emerged as the top performer, showcasing a remarkable accuracy score of 0.9674. This high accuracy underscores the effectiveness of Random Forest in accurately classifying instances into diabetic and non-diabetic categories. Despite its impressive accuracy, Random Forest demonstrated perfect precision, indicating that when it predicted a positive outcome (diabetes), it was highly reliable. However, it exhibited a relatively lower recall compared to k-Nearest Neighbors and Decision Tree algorithms. On the other hand, the Decision Tree algorithm achieved a balanced precision-recall trade-off, suggesting that it was able to correctly identify diabetic cases (high recall) while minimizing false positives (high precision). This balance is crucial in scenarios where both types of errors carry significant consequences, such as in medical diagnosis. Despite Random Forest's superiority in accuracy, Decision Tree's ability to maintain a balanced precision-recall trade-off highlights its competitiveness in diabetes prediction. Additionally, Logistic Regression, while achieving a slightly lower accuracy compared to Random Forest, maintained competitive precision and recall scores. Although Logistic Regression may not have performed as well in terms of accuracy, its ability to provide reliable predictions with balanced precision and recall metrics indicates its suitability as an alternative approach for diabetes prediction.

These findings emphasize the importance of considering multiple evaluation metrics, such as accuracy, precision, and recall, when assessing model performance. While accuracy serves as a fundamental metric, it may not provide a complete picture of a model's effectiveness, especially in the presence of class imbalances.By examining precision and recall alongside accuracy, researchers and practitioners gain a more comprehensive understanding of a model's strengths and weaknesses. In the case of diabetes prediction, where the consequences of misclassification can be severe, prioritizing models with high precision and recall, in addition to accuracy, is crucial for ensuring the reliability and effectiveness of predictive models in clinical practice.

6. Conclusion

In conclusion, leveraging the capabilities of R programming for improving lifestyle choices through diabetes prediction holds substantial promise in enhancing preventive healthcare strategies. Employing machine learning algorithms such as k-Nearest Neighbors, Decision Trees, Random Forests, and Logistic Regression can effectively predict diabetes risk based on various demographic, clinical, and lifestyle factors. The results obtained underscore the importance of accurate risk assessment and personalized intervention strategies in combating the global burden of diabetes. Furthermore, the comprehensive analysis facilitated by R programming not only enables the identification of high-risk individuals but also provides valuable insights for tailored lifestyle recommendations. Continued efforts to refine and optimize these predictive models and integrate them into clinical practice have the potential to empower individuals with actionable insights, ultimately leading to improved health outcomes and a reduction in the prevalence of diabetes-related complications.

References

[1] International Diabetes Federation (IDF), IDF Diabetes Atlas, Ninth Edition, 2021. [Online]. Available: https://diabetesatlas.org/

[2] World Health Organization, Diabetes. [Online]. Available: https://www.who.int/health-topics/diabetes

[3] American Diabetes Association, Understanding Type 2 Diabetes. [Online]. Available: https://diabetes.org/about-diabetes/type-2

[4] Centers for Disease Control and Prevention, Preventing Type 2 Diabetes, 2024. [Online]. Available: https://www.cdc.gov/diabetes/prevent-type-2/index.html

[5] American Diabetes Association Professional Practice Committee, "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022," *Diabetes Care,* vol. 45, pp. S17–S38, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Gregor Stiglic, Roger Watson, and Leona Cilar, "R You Ready? Using the R Programme for Statistical Analysis and Graphics," *Research in Nursing & Health,* vol. 42, no. 6, pp. 494-499, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Hadley Wickham, *Ggplot2: Elegant Graphics for Data Analysis*, Springer International Publishing, pp. 1-260, 2016. [Google Scholar] [Publisher Link]

[8] World Health Organization, Global Report on Diabetes, 2016. [Online]. Available: https://www.who.int/publications/i/item/9789241565257

[9] Narges Razavian et al., "Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors," *Big Data,* vol. 3, no. 4, pp. 277-287, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[10] American Diabetes Association, "Lifestyle Management: Standards of Medical Care in Diabetes—2021," *Diabetes Care,* vol. 44, Supplement 1, pp. S81-S92, 2021. Not found

[11] Paul W. Franks, and Alaitz Poveda, "Lifestyle and Precision Diabetes Medicine: Will Genomics Help Optimise the Prediction, Prevention and Treatment of Type 2 Diabetes through Lifestyle Therapy?," *Diabetologia,* vol. 60, no. 5, pp. 784–792, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[12] George Shaw Jr, Tarun Sharma, and Sthanu Ramakrishnan, "Exploring Diabetes and Users' Lifestyle Choices in Twitter to Improve Health Outcomes," *Proceedings of the Southern Association for Information Systems Conference,* Simons Island, Georgia, USA, pp. 1-6, 2019. [Google Scholar]

[13] Nice Maylani Asril et al., "Predicting Healthy Lifestyle Behaviours among Patients with Type 2 Diabetes in Rural Bali, Indonesia," *Clinical Medicine Insights: Endocrinology and Diabetes,* vol. 13, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Wenwen Du et al., "Thirty-Year Urbanization Trajectories and Obesity in Modernizing China," *International Journal of Environmental Research and Public Health,* vol. 19, no. 4, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Asish Satpathy, and Satyajit Behari, "Machine Learning Prediction of Chronic Diabetes Based on a Person's Demography and Lifestyle Information," *International Journal of Data Science,* vol. 7, no. 3, pp. 210-228, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Surabhi Kaul, and Yogesh Kumar, "Artificial Intelligence-based Learning Techniques for Diabetes Prediction: Challenges and Systematic Review," *SN Computer Science,* vol. 1, no. 6, pp. 1-7, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Quan Zou et al., "Predicting Diabetes Mellitus with Machine Learning Techniques," *Frontiers in Genetics,* vol. 9, pp. 1-10, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[18] Carla K. Miller et al., "Development and Pilot Testing of a Novel Behavioral Intervention for Adults with Type 2 Diabetes Using Intervention Mapping," *Health Psychology and Behavioral Medicine,* vol. 5, no. 1, pp. 317-336, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[19] George Shaw Jr, Tarun Sharma, and Sthanu Ramakrishnan, "Exploring Diabetes and Users' Lifestyle Choices in Digital Spaces to Improve Health Outcomes," *SAIS 2019 Proceedings,* pp. 1-7, 2019. [Google Scholar] [Publisher Link]

[20] Irene Dankwa-Mullan et al., "Transforming Diabetes Care through Artificial Intelligence: The Future is Here," *Population Health Management,* vol. 22, no. 3, pp. 229-242, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[21] Aishwariya Dutta et al., "Early Prediction of Diabetes Using an Ensemble of Machine Learning Models," *International Journal of Environmental Research and Public Health,* vol. 19, no. 19, pp. 1-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Kevin Plis et al., "A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management," *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 35-39, 2014. [Google Scholar] [Publisher Link]

[23] Ann Smith Barnes, "The Epidemic of Obesity and Diabetes," *The Texas Heart Institute Journal,* vol. 38, no. 2, pp. 142-144, 2011. [Google Scholar] [Publisher Link]

[24] American Diabetes Association, What Is the A1C Test?. [Online]. Available: https://diabetes.org/about-diabetes/a1c