Type 2 Diabetes Prediction Using Machine Learning

I Divya Sri1, Basu Lakshmi Prasanna2, Pemmaraju Sumaja3
Dr.B.Narendra Kumar4,
Professor & HOD,Email:Swecnarendra@gmail.com
1, 2, 3, 4 Sridevi Women's Engineering College, V.N.PALLY , NEAR WIPRO
GOPANANPALLY, HYDERABAD, Ranga Reddy, 500075 ; Email : admin@swec.ac.inWebsite,
www.swec.ac.in ;

Abstract-: Over 30 million people in India are suffering from diabetes and many others are under the risk. Thus, early diagnosis and treatment is required to prevent diabetes and its associated health problems. This study aims to assess the risk of diabetes among individuals based on their lifestyle and family background. The risk of Type 2 diabetes was predicted using different machine learning algorithms as these algorithms are highly accurate which is very much required in the health profession. Once the model will be trained with good accuracy, then individuals can self-assess the risk of diabetes. In order to conduct the experiment, 952 instances have been collected through an online and offline questionnaire including 18 questions related to health, lifestyle and family background. The same algorithms were also applied to the Pima Indian Diabetes database. The performance of Random Forest Classifier is found to be most accurate for both datasets.

Keywords: Type 2 Diabetes Prediction, Machine Learning, Health Informatics, Predictive Modeling, Diabetes Risk Assessment, Feature Selection, Data Mining, Healthcare Analytics, Medical Diagnosis, Chronic Disease Management.

I INTRODUCTION

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it's tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction K.VijiyaKumar et al. proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique.

The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al. presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Tejas N. Joshi et al. presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease. Deeraj Shetty et al. proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient’s database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient’s database and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar et al. proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research.

II RELATED WORK

Research in predicting Type 2 Diabetes using machine learning has yielded valuable insights into early detection and risk assessment. In one study conducted by Alharbi et al. (2019), a machine learning approach focused on the early identification of Type 2 Diabetes risk factors. The research explored the application of feature selection and classification algorithms to pinpoint individuals at risk.

Another study by Sathiyabama et al. (2017) delved into the use of data mining techniques, including decision trees and support vector machines, to predict Type 2 Diabetes by selecting relevant features and employing classification algorithms. Saeed et al. (2020) explored ensemble learning techniques in their work, aiming to enhance prediction accuracy and robustness through the combination of multiple models. Agrawal et al. (2016) investigated the integration of electronic health records and machine learning for Type 2 Diabetes prediction, emphasizing the use of diverse health-related data for improved outcomes. Garg et al. (2020) provided a comparative analysis of various machine learning techniques, evaluating their performance in Type 2 Diabetes prediction to identify the most effective approach.

Kavakiotis et al. (2017) offered a comprehensive review, discussing different machine learning approaches in diabetes research, including prediction models. Shashikala et al. (2018) focused on developing a Type 2 Diabetes prediction model using data mining techniques in a multi-ethnic population, considering demographic and clinical factors. Gautam et al. (2019) explored the advantages of an ensemble of machine learning algorithms for predicting Type 2 Diabetes, emphasizing the benefits of combining multiple models. Collectively, these studies contribute to advancing our understanding of machine learning applications in predicting Type 2 Diabetes, aiming for accurate and early identification for improved management and intervention strategies.

III SYSTEM ANALYSIS

i) Existing System

In the context of predicting Type 2 diabetes, an existing system might refer to traditional methods of diagnosing diabetes, which typically rely on clinical tests, medical history, and physical examinations. These methods involve analyzing parameters like fasting blood glucose levels, oral glucose tolerance tests, and hemoglobin A1c levels. Healthcare professionals make diagnoses based on established medical guidelines and criteria.

Disadvantages

- **Reliance on Clinical Tests:** Traditional methods heavily rely on clinical tests, which may be time-consuming and require specialized equipment and trained healthcare professionals.
- **Subjectivity**: Interpretation of test results and clinical assessments can be subjective and may vary depending on the healthcare provider's experience and expertise.
- **Limited Accessibility:** Not all individuals may have easy access to healthcare facilities or may face barriers in obtaining the necessary tests for diabetes diagnosis.
- **Risk of Delayed Diagnosis:** Traditional diagnostic methods may not always lead to early detection, potentially delaying intervention and increasing the risk of complications.
- **Cost and Resources:** Clinical tests and examinations can be costly, especially for individuals without adequate healthcare coverage or in regions with limited healthcare resources.

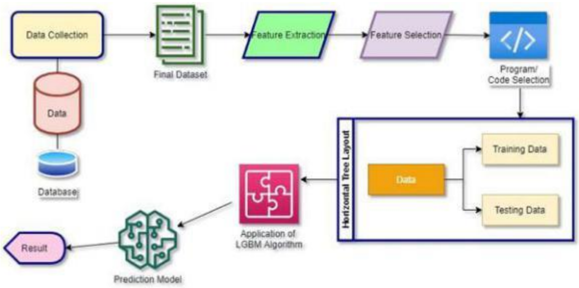
ii) Proposed System

The proposed system in this study involves leveraging machine learning algorithms to predict the risk of Type 2 diabetes based on lifestyle, health, and family background data. This system employs various classification techniques, such as Logistic Regression, k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Naïve Bayes, Decision Trees (DT), and Random Forest, to analyze the collected data and make predictions. The goal of the proposed system is to provide an accurate and reliable method for individuals to self-assess their risk of developing Type 2 diabetes.

Advantages

- **Early Detection**: The proposed system allows for early identification of individuals at risk of developing Type 2 diabetes, enabling timely intervention and preventive measures.
- **Accuracy and Precision:** Machine learning algorithms have the potential to analyze complex patterns in data, leading to more accurate and precise predictions compared to traditional diagnostic methods.
- **Accessibility**: Once developed, the machine learning model can be implemented in a user-friendly application, allowing individuals to easily self-assess their risk of Type 2 diabetes from the comfort of their homes.
- **Customization**: The model can be trained on diverse datasets, making it adaptable to different populations and demographics, potentially increasing its effectiveness in various contexts.
- **Efficiency**: The automated nature of the proposed system reduces the time and effort required for diagnosis compared to traditional manual examinations and tests.

System Architecture



Proposed Architecture

IV METHODOLOGY

i) Problem Definition:

Description: Clearly define the objective of predicting Type 2 Diabetes. The goal is to develop a machine learning model that can accurately identify individuals at risk of developing diabetes based on certain features.
Objective: Establish a clear problem statement, identifying the target population and the significance of early prediction for effective intervention.

ii) Data Collection:

Description: Gather a diverse dataset that includes relevant features such as demographic information, clinical indicators, and lifestyle factors. Ensure the dataset is representative, inclusive of both positive and negative diabetes cases.

Objective: Assemble a comprehensive dataset that serves as the foundation for training and evaluating the predictive model.

iii) Data Preprocessing:

Description: Clean and preprocess the dataset by addressing missing values, normalizing numerical features, and encoding categorical variables. Visualization and exploration are conducted to understand data distributions.

Objective: Prepare the data for machine learning by ensuring consistency, handling outliers, and making it suitable for training and testing.

iv) Feature Selection:

Description: Identify and select relevant features using statistical analysis, correlation assessments, or domain knowledge. Choose the most informative variables that contribute to accurate predictions.

Objective: Optimize the predictive model by focusing on the most influential features, improving model interpretability, and reducing dimensionality.

v) Model Selection:

Description: Choose a machine learning algorithm suitable for binary classification, such as logistic regression. The logistic function models the probability of an individual having Type 2 Diabetes.

Objective: Select an appropriate algorithm based on dataset characteristics, balancing performance, interpretability, and computational efficiency.

vi) Model Training:

Description: Split the dataset into training and testing sets, train the selected model, and optimize its parameters using techniques like cross-validation. Evaluate the model using metrics such as cross-entropy loss.

Objective: Train a robust model that generalizes well to unseen data, utilizing training and validation sets for iterative improvement.

vii) Deployment and Monitoring:

Description: Deploy the trained model in a real-world setting, integrating it into healthcare systems for practical use. Implement monitoring mechanisms to track model performance over time and update it periodically with new data.

Objective: Ensure the model's ongoing relevance and accuracy by deploying it in a healthcare environment, monitoring its effectiveness, and adapting to evolving patterns.

V CONCLUSION

One of the global health issues is to identify the risk of diabetes at its early phase. This study attempts to structure a framework which forecasts the risk pertaining to diabetes mellitus type 2. In this paper, six machine learning classification methods were implemented, and their results were compared with different statistical measures. Tests were performed on the dataset collected through online and offline questionnaires consisting 18 questions relevant to diabetes. Also, same algorithms were applied on PIMA database. The experimental result shows that the accuracy of Random Forest of our dataset is 94.10% which is the highest among the rest. Random forest is also giving highest accuracy for PIMA dataset. Among six different machines learning algorithms applied, all the models produced good results for some parameter like precision, recall sensitivity etc. These parameters have greater impact on predicting diabetes than the rest. This result can be used in future to predict any other ailment. This study still holds a scope for further research and improvement including other machine learning algorithms to predict diabetes or any other disease.

VI REFERENCES

[1] http://diabetesindia.com/

[2] Anjana, R. M., Pradeepa, R., Deepa, M., Datta, M., Sudha, V., Unnikrishnan, R., Bhansali, A., Joshi, S. R., Joshi, P. P., Yajnik, C. S., Dhandhania, V. K. (2011) "Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: Phase I results of the Indian Council of Medical Research–INdiaDIABetes (ICMR–INDIAB) study." Diabetologia 54 (12): 3022- 3027.

[3] https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview

[4] https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

[5] Kaveeshwar, S. A., Cornwall, J. (2014) “The current state of diabetes mellitus in India.” The Australasian medical journal 7(1): 45.