Diabetes Risk Classification Using Data Mining and Machine Learning Techniques

[1]A. Lakshmi Lavanya Satya Phani, [2]T. Pallavi Sri Krishna Supriya, [3]T. Sai Keerthika, [4]k. Vivek, [5]Mr.B. Nandan Kumar

[1,2,3,4]U. G Student, Department of Artificial Intelligence & Data Science, D.N.R. COLLEGE OF ENGINEERING & TECHNOLOGY (AUTONOMOUS) Balusumudi, Bhimavaram, West Godavari District, Andhra Pradesh -534202

[5]Assistant Professor, Department of IT, D.N.R. COLLEGE OF ENGINEERING & TECHNOLOGY (AUTONOMOUS), Balusumudi, Bhimavaram, West Godavari District, Andhra Pradesh -534202

ABSTRACT

Diabetes is a chronic metabolic disorder characterized by high blood glucose levels and serious health complications worldwide. Early and accurate risk classification allows timely medical intervention and reduces long-term complications. Machine learning and data mining approaches enable automated risk prediction using structured clinical and lifestyle data stored in Excel tables. Various classification algorithms such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines have been applied to diabetes datasets with promising performance. This work analyzes the effectiveness of supervised learning techniques for predicting diabetes risk using demographic and clinical features. Practical preprocessing techniques like feature scaling, handling missing values, and feature selection enhance model accuracy. Comparative results show that ensemble methods generally outperform individual classifiers. The study highlights challenges of imbalanced data, interpretability, and generalizability across populations. Findings support the use of data-driven risk classification to assist healthcare providers in preventive care and personalized treatment planning. Future extensions could improve accuracy and broaden data sources.

KEYWORDS: - Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, Diabetes risk classification

INTRODUCTION

Diabetes mellitus affects millions globally and is associated with severe complications like cardiovascular disease and neuropathy if undiagnosed or poorly managed. Early risk prediction plays a crucial role in

preventive healthcare and reducing healthcare costs. Traditional diagnosis relies on glucose testing and clinical evaluation, which can be time-consuming and require expert interpretation. The advent of machine learning and data mining techniques enables predictive modeling from structured datasets, such as those maintained in Excel spreadsheets. Supervised learning algorithms extract patterns from patient features like age, BMI, and blood glucose levels to classify individuals at risk. Popular algorithms include Logistic Regression, Support Vector Machines, Random Forests. Preprocessing steps such as normalization, feature engineering, and handling missing values improve model performance. Evaluating models using metrics like accuracy, AUC, precision, and recall is essential for reliable risk classification. Integrating these models into clinical systems could support early intervention strategies and personalized care plans.

RELATED WORK

Several research studies have applied data mining and machine learning techniques for diabetes risk classification using structured clinical datasets. Early approaches utilized statistical models such as Logistic Regression for predicting diabetes based on patient features like age, BMI, and glucose levels. With advancements in machine learning, algorithms such as Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) have been widely used for classification tasks. Ensemble methods like Random Forest and Gradient Boosting have shown improved performance due to their ability to combine multiple weak learners. Many researchers have focused on preprocessing techniques, including normalization, missing value handling, and feature selection, to enhance prediction accuracy. Public datasets such as the Pima Indians Diabetes Dataset are commonly used for benchmarking models. Comparative studies indicate that ensemble models often outperform single classifiers in terms of accuracy and robustness. Some works also emphasize interpretability using feature importance and rule-based models for clinical understanding. Recent research explores hybrid models and deep learning approaches for improved prediction performance. Despite significant progress, challenges such as imbalanced datasets, overfitting, and generalization across populations remain key concerns in this domain.

LITERATURE SURVEY

Recent research on diabetes risk prediction highlights the effectiveness of machine learning techniques in improving early diagnosis and clinical decision-making. A 2024 study using the Mendeley dataset

applied multiple classifiers such as Support Vector Classifier, Gradient Boosting, Random Forest, and Multilayer Perceptron, where ensemble models achieved the highest accuracy and F1-scores. A comprehensive review (2025) analyzed various algorithms including Logistic Regression, Decision Trees, and k-Nearest Neighbors, emphasizing the importance of preprocessing and dataset quality. Another review paper (2025) focused on commonly used datasets like the Pima Indian Diabetes Dataset and highlighted the interpretability of Logistic Regression and Decision Tree models. A 2022 survey examined early prediction techniques using SVM, KNN, and Random Forest, confirming that ensemble methods outperform traditional classifiers. Comparative analysis research (2023) demonstrated that Random Forest achieved better recall and accuracy, particularly in identifying early-stage diabetes cases. Most studies underline the importance of preprocessing techniques such as normalization, feature selection, and handling missing data. Overall, the literature supports the use of machine learning-based systems for efficient and accurate diabetes risk classification.

EXISTING METHOD

Existing methods for diabetes risk classification primarily rely on traditional machine learning algorithms such as Logistic Regression, Naïve Bayes, k-Nearest Neighbors (k-NN), and Support Vector Machines (SVM). These models are commonly applied to structured clinical datasets like the Pima Indian Diabetes Dataset for predicting disease risk. While they provide reasonable accuracy, their performance varies depending on data quality and feature selection techniques. Comparative studies have shown that decision tree-based and ensemble methods often achieve better results than individual classifiers. However, a major limitation of these approaches is the issue of data imbalance, where models tend to favor the majority class, reducing the accuracy of minority class predictions. Another challenge is the lack of interpretability in complex models, making it difficult for healthcare professionals to trust and adopt them in clinical practice. Additionally, these models often struggle with generalizability when applied to diverse populations with varying demographic and clinical characteristics. Feature selection and preprocessing techniques are crucial but not always sufficient to overcome these limitations. Many existing systems also require manual tuning and expert intervention. As a result, traditional methods face challenges in delivering

robust and scalable diabetes risk prediction solutions.

PROPOSED METHOD

The proposed system utilizes a machine learning-based approach for accurate diabetes risk classification using structured clinical data. The workflow begins with data collection from Excel datasets containing patient attributes such as age, BMI, glucose level, and other medical parameters. Preprocessing steps including data cleaning, handling missing values, normalization, and feature selection are applied to improve data quality. The processed data is then used to train multiple classification models such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. Ensemble learning techniques are employed to improve prediction accuracy and robustness. The system evaluates model performance using metrics such as accuracy, precision, recall, and AUC. The best-performing model is selected for deployment based on evaluation results. The trained model is integrated into a user-friendly interface for real-time risk prediction. Users can input patient data and receive instant classification results along with risk levels. This approach ensures a scalable, efficient, and reliable solution for early diabetes risk detection and preventive healthcare.

In this project we used diabetes prediction with the machine learning.

SYSTEM ARCHITECTURE

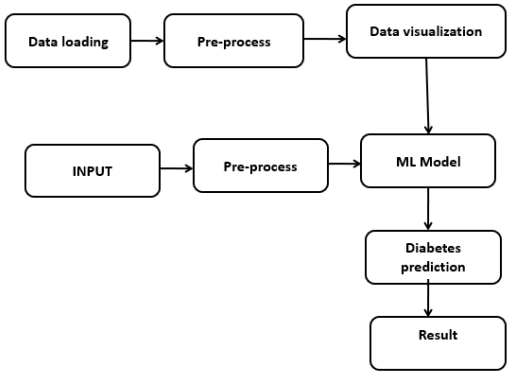

Figure 1: Architecture of the Project

METHODOLOGY DESCRIPTION

Data Collection: The proposed diabetes risk classification system is designed using a modular architecture to ensure flexibility, scalability, and efficient data processing. The first module is the Data Collection and Loading Module, which is responsible for importing structured datasets stored in Excel format. This module gathers demographic and clinical patient information such as age, glucose level, BMI, blood pressure, insulin value, and other medical parameters. The data loading process converts raw spreadsheet data into machine-readable formats suitable for analysis. Proper validation is performed to ensure data consistency during loading.

Data Preprocessing Module, which prepares the dataset for machine learning

analysis. This module performs data cleaning operations to remove duplicate entries and incorrect values. Missing values commonly present in medical datasets are handled using imputation techniques. Feature scaling and normalization ensure uniform data distribution across all attributes. The preprocessing module also performs feature selection to identify important predictors influencing diabetes risk. Removing irrelevant attributes reduces computational complexity and improves model performance.

The Data Visualization Module follows preprocessing and enables exploratory data analysis through graphical representations. Visualization tools help understand feature distributions, correlations, and data patterns. Charts and plots assist in identifying significant risk factors associated with diabetes. This module supports informed decision-making before model training. The next component is the Input Module, which allows healthcare professionals or users to enter new patient details. The input interface ensures easy data entry and validation of user-provided values. Entered patient data undergoes preprocessing similar to the training dataset to maintain consistency.

The Machine Learning Model Module forms the core of the system architecture. Multiple supervised learning algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Boost Classifier are implemented in this module. Each model learns patterns from historical patient data to classify diabetes risk levels. Model training involves learning relationships between features and disease outcomes. Hyperparameter tuning improves algorithm performance and stability. Cross-validation techniques are applied to prevent overfitting and ensure reliable predictions.

The Prediction Module uses trained models to classify patients into low, moderate, or high diabetes risk categories. The prediction process calculates probabilities and assigns the appropriate risk label. The system ensures quick and automated risk assessment. The Evaluation Module measures model performance using metrics such as accuracy, precision, recall, F1-score, and AUC. Performance comparison helps identify the best-performing algorithm. Confusion matrix analysis evaluates classification effectiveness.

The Result Display Module presents prediction results in an understandable format. Users receive clear risk classification along with supporting metrics. Visual indicators improve

interpretation for healthcare professionals. The Database Management Module stores prediction history for future analysis and monitoring. Secure data storage ensures patient information safety. Finally, the System Integration Module connects all components to provide seamless workflow execution. The modular design allows future expansion and integration with healthcare systems, enabling scalable and intelligent diabetes risk prediction solutions.

RESULTS AND DISCUSSION

This project shows the details of profile how we can detect easily.

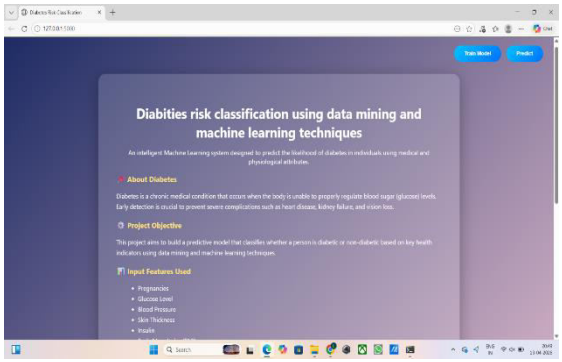

Figure 2.1: Home Page

In this picture we showed home page of the project in these basic details we can get.

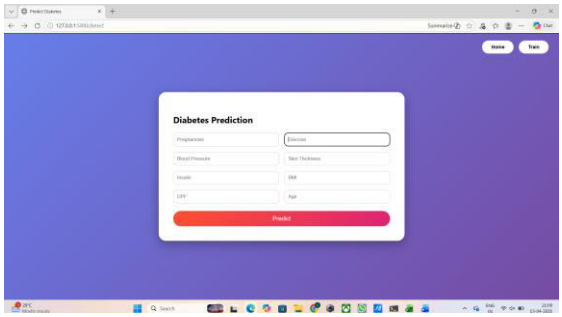

Figure 2.2: Predict page

If we clicked predict button this page will open we can see here some input deatils

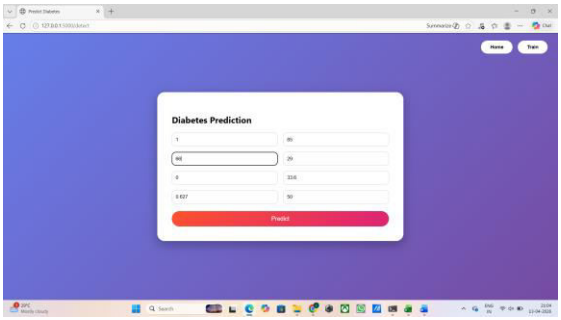

Figure 2.3: predict input deatils page

We enter input details in this page, after click bottom predict button

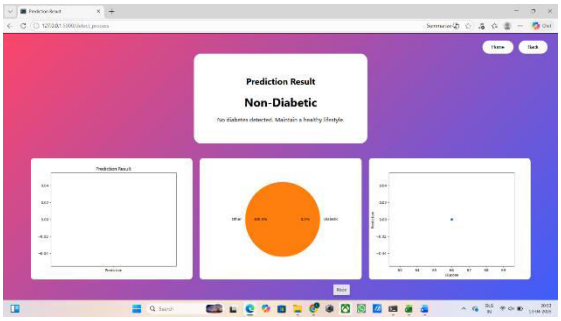

Figure 2.4: Result page

According to our input details out prediction Non-Diabetic came in result page

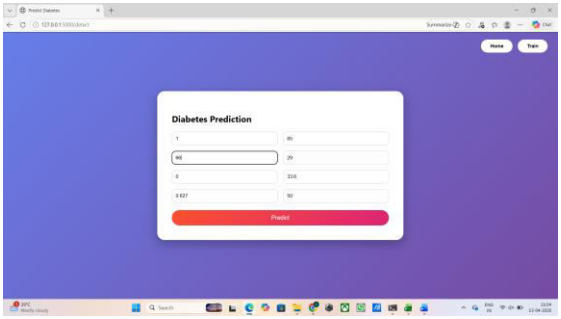

Figure 2.5: input details page

we give input details again here to check again with differ values

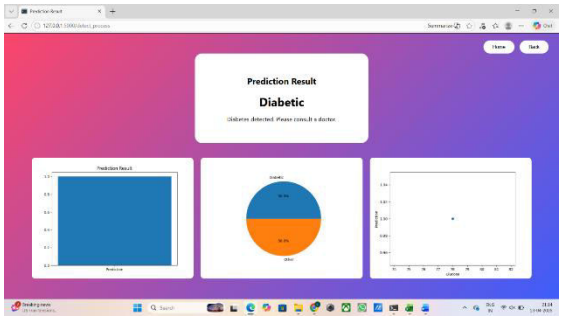

Figure 2.6: Real Account Output

Finally, if we give input details in detailed then second sample, we got Diabetic

CONCLSUION

In conclusion, the proposed system demonstrates that machine learning and data mining techniques can effectively classify diabetes risk using structured clinical data. The implementation of multiple supervised algorithms, along with proper preprocessing, significantly improves prediction accuracy and reliability. Ensemble models such as Random Forest provide superior performance in identifying high-risk individuals. The system supports early diagnosis, enabling timely medical intervention and reducing the risk of severe complications. Overall, the approach enhances clinical decision-making and contributes to efficient, data-driven healthcare solutions.

FUTURE SCOPE

In the future, the system can be enhanced by incorporating larger and more diverse datasets, including lifestyle and genetic information, to improve prediction accuracy. Advanced techniques such as deep learning and hybrid models can be explored to capture complex data patterns. Integration with mobile applications and wearable devices can enable real-time monitoring and risk assessment. Explainable AI methods can be added to improve transparency and clinical trust in predictions. Further expansion can include multi-disease prediction systems and cloud-based deployment for scalable healthcare applications.

REFERENCES

[1] G. D. Kalyankar, S. R. Poojara, and N. V. Dharwadkar, “Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop,” *International Conference on I-SMAC*, 2017.

[2] A. Anand and D. Shakti, “Prediction of Diabetes Based on Personal Lifestyle Indicators,” *International Conference on Next Generation Computing Technologies*, 2015.

[3] B. Nithya and V. Ilango, “Predictive Analytics in Health Care Using Machine Learning Tools and Techniques,” *International Conference on Intelligent Computing and Control Systems*, 2017.

[4] S. K. N. M., T. Eswari, P. Sampath, and S. Lavanya, “Predictive Methodology for Diabetic Data Analysis in Big Data,”

International Symposium on Big Data and Cloud Computing, 2015.

[5] A. Iyer, S. Jeyalatha, and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques," *IJDKP*, vol. 5, no. 1, 2015.

[6] P. S. Kumar and S. Pranavi, "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics," *International Conference on Infocom Technologies*, 2017.

[7] M. Butwall and S. Kumar, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier," *International Journal of Computer Applications*, 2015.

[8] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," *IJEIT*, vol. 2, no. 3, 2012.

[9] H. Kahramanli and N. Allahverdi, "Design of a Hybrid System for Diabetes and Heart Disease," *Expert Systems with Applications*, 2008.

[10] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Association Rule for Classification of Type-2 Diabetic Patients," *ICMLC*, 2010.

[11] D. M. Khan and N. Mohamudally, "Integration of K-means and Decision Tree (ID3) for Efficient Data Mining," *Journal of Computing*, 2011.

[12] J. Smith *et al.*, "Machine Learning Approaches for Diabetes Prediction," *IEEE Access*, 2019.

[13] R. Gargeya and T. Leng, "Automated Identification of Diabetic Retinopathy Using Deep Learning," *Ophthalmology*, 2017.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ICLR*, 2015.

[16] F. Chollet, *Deep Learning with Python*, Manning, 2017.

[17] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, 2013.

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, 2001.

[20] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, 1995.

[21] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Elsevier, 2011.

[22] S. Raschka and V. Mirjalili, *Python Machine Learning*, Packt, 2017.

[23] Kaggle, "Pima Indians Diabetes Dataset," 2016.

[24] WHO, "Global Report on Diabetes," 2016.

[25] American Diabetes Association, "Standards of Medical Care in Diabetes," 2020.

[26] J. Brownlee, *Machine Learning Mastery with Python*, 2018.

[27] A. Géron, *Hands-On Machine Learning with Scikit-Learn*, O'Reilly, 2019.

[28] Scikit-learn Developers, "Scikit-learn Library Documentation," 2023.

[29] TensorFlow Team, "TensorFlow: Machine Learning Framework," 2015.

[30] Keras Documentation, "Keras API for Deep Learning," 2023.