Predicting Diabetes Risk Using Machine Learning Algorithms

Hrishit Madhavi
School of Computer Science and Engineering
MIT World Peace University
Pune ,India
1032220164@mitwpu.edu.in

Hrishikesh Iyer
School of Computer Science and Engineering
MIT World Peace University
Pune ,India
1032220223@mitwpu.edu.in

Ayush Patil
School of Computer Science and Engineering
MIT World Peace University
Pune ,India
1032220274@mitwpu.edu.in

Rujuta Kulkarni
School of Computer Science and Engineering
MIT World Peace University
Pune ,India
1032220271@mitwpu.edu.in

***Abstract*—*In recent years, the incidence of diabetes has increased significantly, leading to a growing need for accurate and efficient risk prediction methods. Machine Learning (ML) algorithms have emerged as powerful tools for predicting diabetes risk based on various factors such as age, lifestyle, and medical history. This paper explores the use of multiple ML algorithms, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM), for predicting the likelihood of diabetes. The models are trained on patient data to identify patterns and risk factors associated with diabetes. Through comparative analysis, we demonstrate the performance of each model, focusing on improving prediction accuracy and minimizing false positives. Our approach aims to provide a reliable and interpretable tool for early diabetes risk assessment, facilitating timely interventions and improving patient outcomes.*

Keywords*—*Diabetes, Machine Learning, Risk Prediction, Logistic Regression, Decision Trees, Random Forest, SVM, Classification.

1. INTRODUCTION

Diabetes mellitus, a chronic disorder characterized by high blood glucose levels, poses significant global health challenges. The World Health Organization (WHO) reports an increasing prevalence of diabetes, leading to substantial morbidity and mortality. Early detection and timely intervention are crucial for effective management and prevention of complications. Traditional diagnostic methods often rely on clinical assessments that may not be timely or comprehensive.Recent advancements in machine learning (ML) provide promising solutions for diabetes prediction and diagnosis. ML algorithms analyze extensive datasets to uncover hidden patterns, facilitating accurate predictions of disease onset. Various studies have utilized supervised and unsupervised learning techniques with diverse datasets to predict diabetes risk effectively.This report offers an overview of diabetes prediction using ML algorithms, focusing on methodologies, datasets, and algorithm performance, aiming to enhance diagnostic precision and improve patient outcomes.

2. LITERATURE SURVEY

In this section, some closely related works are discussed briefly.

Ameer Ali, Mohammed Alrubei, Laith Falah Mohammed Hassan,Mohannad Al-Ja'afari And Saif Abdulwahed used the dataset that was generated by the criteria of the American diabetes association and 4900 samples were used for training stage; 100 samples used for testing. The result show that the KNN types (Fine, Weighted, Medium, Cubic) give high accuracy. Fine KNN is considered the most suitable.

Krati Saxena, Dr. Zubair Khan, Shefali Singh used K Nearest Neighbor Algorithm for the diagnosis of diabetes mellitus. Accuracy and error rates have been calculated for K = 3 and K = 5 where 70% and 69% accuracy were achieved respectively.

Huma Naz and Sachin Ahuja used the PIMA Indian dataset. Deep Learning was the presented model; resulting in accuracy of 98.07%.

Kamrul Hasan, Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan proposed a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were employed. The weighted ensembling of different ML models is also proposed, in this literature, to improve the prediction of diabetes where the weights are estimated from the corresponding Area Under ROC Curve (AUC) of the ML model.

Isfafuzzaman Tasin, Tansin Ullah, Nabil Sanjida and Islam Riasat Khan worked on a private dataset of female patients in Bangladesh along with the PIMA dataset. problem. The authors used machine learning classification methods, that is, decision tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques. models, the proposed system provided the best result in the XGBoost classifier with the ADASYN approach with 81% accuracy.

Deepti Sisodia, Dilip Singh Sisodia used three machine learning classification algorithms: Decision Tree, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. Naive bayes had the highest accuracy of 76%.

Mitushi Soni and Dr. Sunita Varma planned to use are K-Nearest Neighbor, Logistic Regression, Decision Tree, Support Vector Machine, Gradient Boosting and Random Forest. Random Forest was observed to perform the best with 77% accuracy.

Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjeeb used adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors. This classification is done across three different ordinal adults' groups in Canadian Primary Care Sentinel Surveillance network. Evaluation of results indicated that adaboost ensemble method outperforms other techniques.

Aishwarya Mujumdar, Dr. Vaidehi V performed a comparative analysis of different ML models on PIMA Diabetes Dataset and a private dataset. The Evaluation showed that AdaBoost Classifier with application of has an accuracy of 98.8% on the private dataset; while it has 77% accuracy on PIMA Dataset.

3. RESEARCH METHODOLOGY

We have followed the below steps to effect this research:

a. Dataset Selection

We chose the **PIMA Indian Diabetes Dataset (PIDD)** for this project, which is available through repositories like UCI and Kaggle. This dataset is widely used in medical research for predicting diabetes and includes health records of Pima Indian women aged 21 years and older. The PIDD is ideal for diabetes prediction due to its comprehensive set of features, which include medical parameters such as glucose concentration, BMI, insulin levels, and age. The dataset consists of 768 records with 8 input features and 1 binary output label (1: diabetic, 0: non-diabetic). Some features of this dataset were :

- Number of pregnancies
- Plasma glucose concentration
- Diastolic blood pressure
- Triceps skinfold thickness
- 2-hour serum insulin
- BMI
- Diabetes pedigree function
- Age

b. Data Cleaning and Preprocessing

To ensure the quality and integrity of the dataset, we performed several data cleaning and preprocessing steps:

1. Handling Missing Values: Some features in the dataset, such as BMI, glucose concentration, and insulin levels, contained missing or impossible values (e.g., 0 for BMI or glucose). Instead of discarding these rows, we applied regression-based imputation techniques:
    - Linear Regression Imputation: For continuous variables like glucose concentration and BMI, we employed linear regression models to predict missing values based on other correlated features.
    - Logistic Regression Imputation: For binary variables and to estimate missing categorical values, logistic regression was used to fill in missing data points.

2. Data Standardisation: To ensure that all features are on the same scale, we standardised continuous variables like glucose concentration, BMI, and insulin levels using Z-score normalisation. This process helps improve the performance of distance-based algorithms such as KNN.

3. Splitting the Dataset: After preprocessing, the dataset was split into two parts: 80% for training and

20% for testing. This division ensures that our model can be trained on a majority of the data while being evaluated on an unseen subset.

c. Model Implementation

1. K-Nearest Neighbours (KNN): The KNN algorithm was chosen for its simplicity and effectiveness in classification problems. It classifies a new data point based on the majority class of its K-nearest neighbours in the feature space.

2. Logistic Regression: Logistic Regression was used for its robustness in binary classification tasks. It provides a probabilistic framework and estimates the probability of diabetes based on the logistic function.

3. Random Forest: We employed Random Forest for its ability to handle non-linear relationships and its robustness against overfitting. The model builds multiple decision trees and aggregates their results to improve prediction accuracy.

4. Decision Tree: The Decision Tree algorithm was chosen due to its interpretability. It builds a tree-like structure where each internal node represents a feature, and each leaf node represents the output class (diabetes or non-diabetes).

d. Model Training and Evaluation

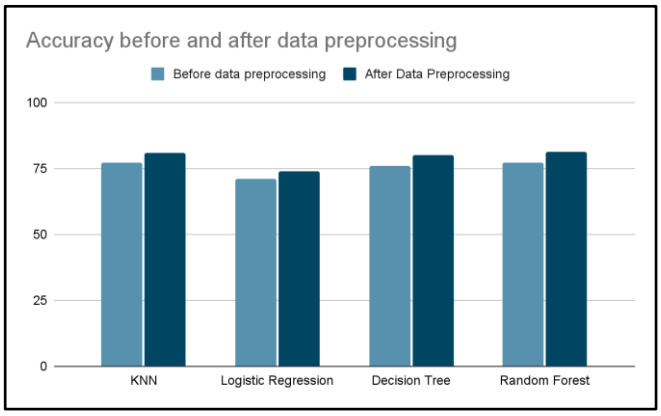
Each model was trained on the preprocessed training data, and hyper-parameter tuning was performed using grid search and cross-validation to optimise performance. After training, we evaluated the models on the test dataset using several metrics, including:

- Accuracy
- Precision
- Recall
- F1-score

The results of each model were compared to determine the most effective algorithm for predicting diabetes.

4. RESULTS

Model	Accuracy	Limitations
KNN	80	Slow for large datasets due to high computation.
Logistic Regression	74.04	Struggles with non-linear data.
Decision Tree	80	Prone to overfitting small datasets.
Randon Forest	80.1	Can overfit and is slow with many trees.



REFERENCES

[1] Ameer Ali, Mohammed Alrubei, Laith Falah Mohammed Hassan,Mohannad Al-Ja'afari And Saif Abdulwahed. (2020). Diabetes Classification Based on KNN. IIUM Engineering Journal, Vol. 21, No. 1, 2020. https://doi.org/10.31436/iiumej.v21i1.1206.

[2] Krati Saxena, Dr. Zubair Khan, Shefali Singh. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbour Algorithm. International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 4, July-Aug 2014.

[3] Huma Naz & Sachin Ahuja. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & MetabolicDisorders(2020).https://doi.org/10.1007/s40200-020-00520-5

[4] Kamrul Hasan, Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE. Digital Object Identifier 10.1109/ACCESS.2020.2989857.

[5] Isfafuzzaman Tasin, Tansin Ullah, Nabil Sanjida and Islam Riasat Khan. (2022). Diabetes Prediction Using Machine Learning and Explainable AI Techniques. Healthcare Technology Letters. DOI/ 10.1049/htl2.12039.

[6] Deepti Sisodia, Dilip Singh Sisodia. (2018). Diabetes Prediction using Machine Learning Algorithm. Procedia Computer Science 132 (2018) 1578–1585.

[7] Mitushi Soni and Dr. Sunita Varma. (2020). Diabetes Prediction using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT) International Journal of Engineering Research & Technology (IJERT).

[8] Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

[9] Aishwarya Mujumdar, Dr. Vaidehi V. (2019). Diabetes Prediction using Machine Learning Algorithms. INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019. Procedia Computer Science 165 (2019) 292–299