PROGNOSIS FOR DIABETES USING: MACHINE LEARNING

[1]Tulsi Bhalani, [2]Ami Mehta

[1]PG Scholar, Computer Science and Engineering, Dr. Subhash University, Gujarat, India
[2]Assistant Professor, Computer Engineering, Dr. Subhash University, Gujarat, India

---***---

Abstract - *Diabetes Mellitus, a chronic metabolic disorder affecting millions worldwide, presents substantial challenges to both healthcare systems and individual human beings. In this study the use of machine learning algorithms to prognosis diabetes by focusing on long term outcomes such as disease progression, complications etc. by using the dataset of diabetes patients records, various machine learning models are trained and evaluated like as SVM, Naïve Bayes, decision tree, random forest, logistic regression, gradient boosting etc. for compaction of treatment strategies and to get improved patient outcome. The highest accuracy is obtained from the hybrid of Correlation Feature Selection technique with XGB Algorithm that is 89%.*

***Key Words*: Diabetes Mellitus, Prognosis, Chronic metabolic disorder, Machine learning, Disease progression, Support Vector Machine (SVM), Decision tree, Random forest, Logistic regression, Gradient boosting, Patient outcomes, XGB Algorithm, Hybrid model, Healthcare, Correlation Feature Selection**

1.INTRODUCTION

Diabetes mellitus is a pervasive and chronic health condition that affects millions of individuals worldwide. Characterized by high blood glucose levels, diabetes can lead to severe complications, including cardiovascular disease, nerve damage, and kidney failure, if not managed properly. The World Health Organization estimates that the prevalence of diabetes has been steadily increasing, underscoring the urgent need for effective management and intervention strategies.

A critical aspect of managing diabetes is the ability to accurately predict its progression. Early prognosis allows for timely intervention, potentially preventing severe complications and improving the quality of life for patients. However, traditional prognostic methods often fall short due to the complex and multifactorial nature of the disease.

In recent years, machine learning has emerged as a powerful tool in the field of healthcare, offering advanced techniques for analyzing large datasets and uncovering patterns that may not be apparent through conventional methods. Machine learning algorithms can process vast amounts of patient data, including medical history, lifestyle factors, and genetic information, to predict disease outcomes with high accuracy.

This study aims to leverage machine learning to enhance the prognosis of diabetes. By employing supervised Learning algorithms, we seek to develop a predictive model that can identify patients at high risk of disease progression and suggest personalized intervention strategies. The following sections will delve into the methodology, dataset, and specific machine learning techniques used, followed by an analysis of the results and their implications for clinical practice.

1.1 Problem Statement:

Diabetes prognostic models are usually detached treating patients as homogeneous entities within the groups that are uniform. Machine learning algorithms have the potential to identify specific variations in health parameters and lifestyle, leading to more customized and effective prognostic models. A more accurate and comprehensive prognosis is made feasible by the integration of genetic information, wearable data, and electronic health records in healthcare. This method provides an in-depth view of an individual's health. The analysis of long-term information is made possible by the use of machine learning algorithms, whereas traditional models may not be able to fully capture temporal patterns. This method contributes to the identification of patterns and trends, resulting in a more thorough prognosis.

1.2 Need of Research:

According to the research India ranks second behind China in Diabetes. The number of people suffering from Diabetes is increasing, there is continuous change in the statistics and frequent updates are required. The cause of Diabetes between the ages of 20-79 years in India is recorded 8.9%. As per the current research Diabetes cases in adults are expected to reach 100 million by 2030. Currently 1 in 6 adults are recorded with diabetes.

As the statistics are increasing rapidly we have proposed the research to predict the occurrence of diabetes by analyzing the symptoms such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin and many more.

Here we use machine learning for prognosing Diabetes. There are mainly three types of machine learning algorithms like Supervised Learning, Unsupervised Learning and Reinforcement Learning. Here we have taken the labeled dataset so that we are using the supervised learning algorithm. There are many algorithms such as Naïve Bayes, Support Vector Machine, Decision Tree, KNN etc.

1.3 Scope:

Implement a feedback mechanism to enhance machine learning models' performance. Tailor models to individual patient characteristics, considering genetic, lifestyle, and environmental factors. Explore integrating diverse data sources like genetic information, health records, and lifestyle information for a comprehensive understanding of diabetes risk. Machine learning models can be used to identify trends, allocate resources efficiently, and design targeted public health interventions for diabetes. Collaboration between researchers, healthcare professionals, data scientists, and technology experts is encouraged for a comprehensive approach. Long-term studies are conducted to assess the effectiveness of these models, ensuring accuracy and adaptability to individual health changes. Ethical concerns related to data privacy and bias are addressed, and guidelines are developed to ensure ethical deployment. Cost-effectiveness and economic impact of early intervention and prevention strategies are also evaluated.

2. RELATED WORK

According to [1] the researcher have used AHDHS Stacking Methodology which generates 93.09% Accuracy. They have used live dataset for input.

By concluding they have said AHSDHS Stacking algorithm can be used for both T1DM and T2DM and can be used for other diseases too, similarly [2] the Researcher have used methodology Multilayer Perceptron, Support Vector Machine, Random Forest with that the accuracy generated is SVM 96% MLP 98% . They have conclude by saying More balanced Dataset should be used to get better Accuracy.

According to [3] LSTM, BILSTM, Feed Forward Neural Network methodology used and the highest Accuracy is Generated by Random Forest the highest accuracy of 82.26%. The research gap says that This model can be applied to the real patient dataset.

In [4] Random Forest, Decision Tree, KNN, Naïve Bayes Methodologies are used from that the highest Accuracy is Generated by Random Forest with Accuracy 92.6%. In future work To compete with state-of-the-art models achieving above 90% accuracy, the implementation of ensemble models should be utilized.

In [5] Random Forest, Decision Tree, KNN, and Naïve Bayes methodologies the generated Accuracy is 75%. The researchers have concluded that Linear Regression with ensemble techniques should be used to increase the accuracy. In [6]Naive Bayes, Decision Trees, Random Forests, and Logistic Regression the Accuracy generated is 78% the researcher concluded that Comparison of state-of-the-art techniques such as XGBoost, AdaBoost, or high-layer DNNs would probably provide better insights regarding the predictive limitations of the constructed dataset.

3. PROPOSED METHODOLOGY

The purpose of the proposed research is to prognosis if the patient is diabetic or not, based on features like Pregnancy, BMI, Insulin, Diabetes Pedigree Function, Age, and Outcomes by using a Support Vector Machine(SVM), Linear Regression, Decision Tree, KNN, Random Forest, XG-Boosting. The goal of this Research is to increase accuracy of Prognosis of diabetes.

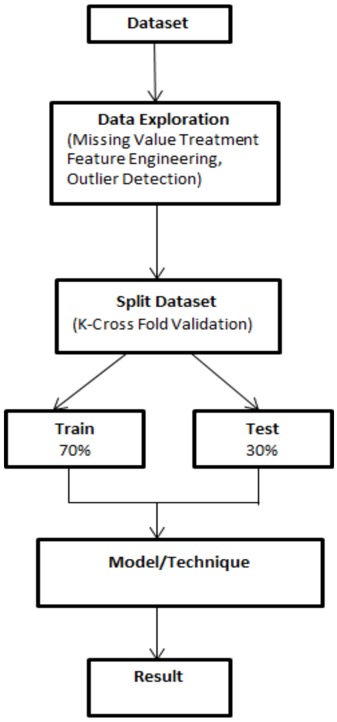

Fig.1 Proposed Framework

An accurate, multidisciplinary approach combining knowledge from data science, machine learning, and medical science is required when researching the prognosis of diabetes using machine learning. In this regard, it is necessary to consider several important aspects:

1. Gathering Data:

Compiling comprehensive patient data entails obtaining data on clinical measurements such as BMI, HbA1c, and blood glucose levels, as well as demographics, medical history, and lifestyle factors.

Furthermore, the application of biosensors and wearables for tracking physical activity levels and continuous glucose monitoring (CGM) can be investigated for the real-time monitoring of physiological data.

2. Pre-processing Data:

For data cleaning, it involves handling missing values, outliers, and inconsistencies within the dataset.

Here in the Missing Value Treatment the entire Dataset is searched and finds the missing value.

Outlier is used to detect the contrast data from the entire Dataset.

To observe, the Outlier Interquartile Range (IQR) is used.

The local Outlier Factor (LOF) method is used in anomaly detection to identify outliers in a dataset.

3. Feature Selection:

In Feature Selection, two techniques are used which are the Information Gain and Correlation matrix Method.

In Information Gain, the most relevant features for a given task are selected, which measures the effectiveness of a feature in classifying or predicting a target variable by quantifying the amount of information gained about the target variable when the feature is known.

The Graph for Prognosing the Diabetes with Information Gain Technique is as:

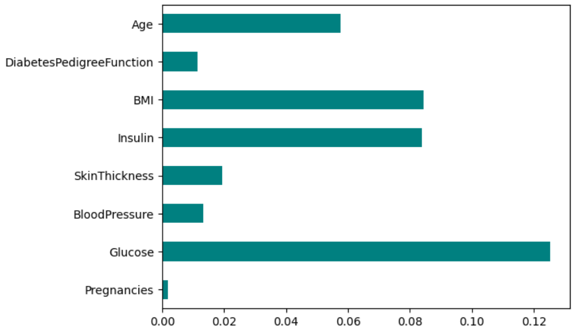

Fig. 2 Feature Selection Using Information Gain

Here Features like Glucose, BMI, Insulin, and Age are selected which are processed with the Outcome Feature.

Correlation-based feature selection is a technique used to identify relevant features for prognosis. The basic idea is to measure the correlation between each feature of the Dataset and the target variable, and select the features with the highest correlation.

The Graph for Prognosing the Diabetes with correlation Technique is as:

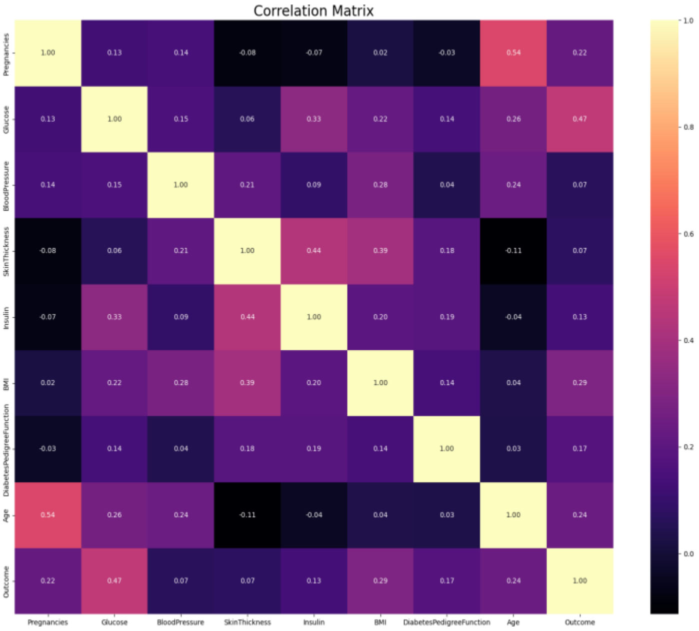

Fig. 3 Feature Selection Using Correlation

Here Features like Glucose, BMI, and Insulin are selected which are processed with the Outcome Feature.

4. Splitting Dataset:

The Dataset is Split in the ratio of 70%-30% into Training Dataset and Testing Dataset

The Selected Features from the Feature Selection Techniques are given as input to the Training Dataset; the Outcome Feature is given input as a Testing Dataset. Splitting of Dataset is done by Cross-Validation Technique.

5. Model Evaluation:

For Evaluating the Result various Models are tested like Logistic Regression, Gradient-Boosting, SVM, Decision Tree, KNN, and Random Forest with Hybrid of Feature Selection Techniques like Information Gain and Correlation Matrix.

The Result table is as follows:

Models	Accuracy (Information Gain)	Accuracy(Correlation)
LR	0.769737 (0.048256)	0.848684 (0.036866)
KNN	0.761842 (0.039671	0.840789 (0.023866)
CART	0.702632 (0.064782)	0.857895 (0.024826)
RF	0.756579 (0.048256)	0.881579 (0.026316)
SVM	0.769737 (0.045675)	0.853947 (0.036488)
XGB	0.753947 (0.054585)	0.890789 (0.020427)

[Comparisons of Feature Selection Technique with Correlation Technique]

6. Evaluation Matrices:

To assess the performance of the model, select appropriate evaluation metrics such as accuracy. Assess the clinical relevance of these measures to make sure they support the goals of diabetes prognosis.

Accuracy

Accuracy=Total Number of PredictionsNumber of Correct Predictions×100%

In this equation:

"Number of Correct Predictions" refers to the number of predictions made by a model that match the actual outcomes.

"Total Number of Predictions" is the total number of predictions made by the model.

4. CONCLUSIONS AND FUTURE WORK

As Diabetes patients are rapidly increasing day to day it is very crucial to predict diabetes according to its features like BMI, Glucose, Skin- Thickness, Diabetes Pedigree Function, Pregnancies, etc. to get proper treatment and control before it causes, predict diabetes at early stage various Machine Learning Techniques are used.

From the dataset which is referred from the Kaggle two Feature Selection Techniques Information Gain and Correlation Methods are used for the Feature Selection, by using Information Gain the selected Features are: Glucose, BMI, Insulin, and Age. By using Correlation Method the features like Glucose, BMI, and Insulin are selected, and the features from individual methods are given as input to the Various Machine Learning Algorithms like Logistic Regression, Random Forest, Decision Tree, SVM, Extreme Gradient Boosting, KNN and compared by using these two Feature Selection techniques. Better results are generated from the Correlation Method hybrids with various Machine Learning Algorithms mentioned.

In the future one can extend the work by using an accurate and live dataset with the hybrid correlation Technique with MultiLayer Perceptron.

REFERENCES

[1] Zaiheng Zhang. “A novel evolutionary ensemble prediction model using harmony search and.” *Elseiver*, vol. 100605, 2023, p. 3.

[2] MaximillianTanea, and Matthew. “Using Machine Learning for the Prediction of Diabetes.” *ELSSEVIER*, vol. 20, 2020.

[3] PriyankaRajendra. “Prediction of diabetes using logistic regression and ensemble techniques.” *ELSSEVIER*, vol. 1, 2020.

[4] FAZAKIS, NIKOS. “Machine Learning Tools for Long-Term Type 2.” *IEEE*, vol. 2021.

[5] SaúlLangaricaa, c. “A meta-learning approach to personalized blood glucose prediction in type 1.” *ELSSEVIER*, vol. 2, 2021.

[6] Victor Chang a, ∗. “An assessment of machine learning models and algorithms for early.” *ELSSEVIER*, 2021.

[7] A. Z. Woldaregay, E. Årsand, S. Walderhaug. “Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes.” vol. 98, 2019.

[8] Y.-F. Du, H.-Y. Ou, E. A. Beverly, and C.-J. Chiu. “Achieving glycemic control in elderly patients with type 2 diabetes: A critical comparison of current options.” vol. 9, 2014.

[9] Y. Gao, Y. Xiao, R. Miao, J. Zhao, M. Cui, G. Huang, and M. Fei. “The prevalence of mild cognitive impairment with type 2 diabetes mellitus among elderly people in China: A cross-sectional study.” vol. 62, 2016.

[10]D. “Diagnosis and classi_cation of diabetes mellitus,'' Diabetes care. Mellitus.” vol. 28, 2015.

[11]M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan. "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers." *IEEE*, vol. 8, 2020.

[12]Ilango, B. Nithya and Dr. V. ``Predictive Analytics in HealthCare Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems." vol. 7, 2017. +

[13]Classification and diagnosis of diabetes: Standards of medical care in diabetes_2020." Association, American Diabetes, vol. 43, 2020.

[14]Ahuja, H. Naz and S. "Deep learning approach for diabetes prediction using PIMA Indian dataset." J. Diabetes Metabolic Disorders, vol. 19, 2020.

[15]Y. Liu, Q. Wang, K. Wu, Z. Sun, Z. Tang, X. Li, B. Zhang. "Anthocyanins' effects on diabetes mellitus and islet transplantation." 2022.