Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(4), 2034-2042 | ISSN:2251-6727

A Study on Data Analysis and Prediction of Diabetes Using Machine Learning Models

Dr. N. Ravi[1], Dr. P. Rajesh[2]

[1]Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram – 608 102, Tamil Nadu, India.

[2]Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram – 608 102, (Deputed from Dept. of Computer and Information Science, Annamalai University, Annamalainagar-608 002) Tamil Nadu, India.

(Received: 02 September 2023 Revised: 14 October Accepted: 07 November)

KEYWORDS

Machine learning, diabetes prediction, decision tree, correlation coefficient, and test statistics.

ABSTRACT:

It's crucial to emphasize that although these approaches may aid in diabetes prediction, it is essential for a healthcare provider to make the diagnosis and offer guidance on managing the condition. While AI and machine learning-based diabetes prediction models are advancing in accuracy and sophistication, they should complement medical expertise as supportive tools. Data mining is typically described as the practice of employing computer systems and automation to explore extensive datasets, identifying patterns and trends, and converting these discoveries into valuable business insights and predictive analyses. This paper considers diabetes prediction-related dataset data like gender, age, hypertension, heart disease, smoking history, bmi, HbA1c level, Blood Glucose level, diabetes. The machine learning approaches which is used to analysis and predict the dataset using linear regression, decision stump, M5P, random forest, random tree, and REP tree**.** Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

1. Introduction and Literature Review

The realm of diabetes prediction research, leveraging data mining and machine learning, stands as a compelling and essential area of investigation. This field employs advanced computational methods to scrutinize vast datasets encompassing factors associated with diabetes risk, patient characteristics, and health metrics.

Machine learning engineers craft autonomous systems through programming. These algorithms automatically gather data and leverage it to acquire knowledge. These systems are anticipated to discern data patterns and autonomously make significant decisions. Data mining is employed for delving into ever-expanding databases and enhancing market segmentation. By scrutinizing the connections between variables like customer age, gender, preferences, and more, it becomes feasible to predict their behaviors and tailor personalized loyalty campaigns accordingly.

The objective of this current study is to systematically review the utilization of machine learning, data mining techniques, and tools in the domain of diabetes research, focusing on a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and d) Health Care and Management. Notably, the primary category appears to be Prediction and Diagnosis. A diverse range of machine learning algorithms was employed, with approximately 85% utilizing supervised learning approaches and 15% employing unsupervised methods, specifically association rules. Notably, Support Vector Machines (SVM) emerged as the most prominent and successful algorithm. Clinical datasets were the predominant data type utilized in the selected articles, highlighting the potential for extracting valuable insights to advance our understanding and investigation of Diabetes Mellitus [1].

Accurate prediction is crucial for diagnosing Diabetes, and data mining plays a pivotal role in extracting meaningful information from vast datasets. The primary aim of data mining is to uncover novel patterns and offer valuable insights to aid in medical diagnosis and treatment. This project seeks to mine Diabetes data to enhance classification efficiency by exploring various data mining methods and techniques, ultimately identifying the most suitable approaches for Diabetes dataset classification and pattern extraction [2].

Chronic diabetes care generates extensive data related to self-management and clinical oversight. This paper presents a dual perspective on this information. First, it introduces a predictive model for short-term glucose regulation based on machine learning, with the goal of preventing daily hypoglycemic events and prolonged hyperglycemia. Second, it proposes data mining approaches for understanding and forecasting long-term glucose control and the incidence of diabetic complications [3].

The global increase in diabetic patients, as reported by WHO, necessitates early identification of diabetes cases. Data mining plays a pivotal role in diabetes research, unearthing hidden insights within vast diabetes-related datasets. Various data mining techniques contribute to improving diabetes research and healthcare for patients. This paper provides an overview of common data mining methods applied to diabetes data analysis and disease prediction [4].

Clinical decision-making relies on available information to guide physicians. Data mining methods have become crucial in medical research, especially in analyzing large volumes of medical data. This study employs data mining techniques, including Fuzzy C-Means (FCM) and Support Vector Machines (SVM), to analyze a dataset related to diabetes diagnosis. This dataset comprises 9 input attributes related to clinical diabetes diagnosis and one output attribute indicating the presence of diabetes in 768 cases [5].

Data mining is a valuable tool for extracting hidden knowledge from extensive databases. In this paper, the dataset focuses on weather conditions for specific days, with attributes such as Outlook, Temperature, Humidity, Windy, and the binary class "Play Golf." Seven classification algorithms, including J48, Random Tree, Decision Stump, Logistic Model Tree, Hoeffding Tree, Reduce Error Pruning, and Random Forest, are utilized. Among these, the Random Tree algorithm yields the highest accuracy of 85.714% [6].

Diabetes Mellitus, characterized by high blood sugar levels, can result from insufficient insulin production or improper cellular responses to insulin. This study aims to develop a data mining model for predicting suitable dosage plans for diabetes patients. Using medical records from 89 patients, 318 diabetes assays are extracted. ANFIS and Rough Set methods are employed for dosage planning, with ANFIS proving more successful and reliable than the Rough Set method [7].

Data mining techniques are powerful tools for knowledge extraction, particularly in medicine. This paper investigates the application of data mining in diabetes self-management (DSM) through a systematic mapping study. The study analyzes the years and sources of DSM publications, the types of diabetes studied, the most common data mining tasks and techniques, and the considered functionalities. Notably, prediction tasks and Neural Networks are frequently employed, with a focus on Type 1 Diabetes Mellitus [8].

Data mining techniques are explored for early prediction of diabetes, a chronic disease affecting multiple organs. A dataset of 768 instances from the PIMA Indian Dataset is used to assess the accuracy of data mining techniques in prediction. The analysis demonstrates that the Modified J48 Classifier yields the highest accuracy among the tested techniques [9].

Data mining is a valuable tool for extracting concealed information and making predictions based on stochastic sensing. This paper introduces an efficient assessment of groundwater levels, rainfall, population, food grains, and enterprises data through stochastic modeling and data mining. The study includes novel data assimilation analysis to predict groundwater levels effectively [10] and [11].

This study involves data related to chronic diseases, with attributes including topics, questions, data values, low confidence limits, and high confidence limits. Five classification algorithms are used for training and testing. Among them, the M5P decision tree algorithm is found to be the most effective in building the model compared to other decision tree approaches [12].

2. Backgrounds and Methodologies

A data mining decision tree is a widely used machine learning technique for classification and regression tasks. It visually depicts a sequence of decisions and their possible outcomes in a tree-like structure. Each internal node represents a decision based on a specific feature, and each branch corresponds to the potential result of that decision. The tree's leaf nodes represent the final decision or the predicted outcome. The "CART" (Classification and Regression Trees) algorithm is the most used algorithm for building decision trees [13].

2.1 Linear Regression

Linear regression is a statistical technique employed to comprehend and forecast the connection between two variables by discovering the optimal straight line that most effectively aligns with the data points. It aids in ascertaining how alterations in one variable correspond to changes in another, proving valuable for predictions and trend recognition.

The core idea of linear regression is to find the best-fitting straight line (also called the "regression line")

through a scatterplot of data points. This line represents a linear equation of the form:

$$y = m_x+b \qquad \dots (1)$$

Where:

- y is the dependent variable (the one you want to predict or explain).
- x is the independent variable (the one you're using to make predictions or explanations).
- m is the slope of the line, representing how much
- y changes for a unit change in x.

b is the y-intercept, indicating the value of y when x is 0.

2.2 Decision Stump

A decision stump is a straightforward machine-learning model for binary classification tasks. It is the most basic form of a decision tree with a single level or depth of one. In a decision stump, only one feature (attribute) of the data is used to decide, and it splits the data into two subsets based on a threshold value for that feature. The decision stump can be understood as choosing one part, choosing a threshold, assigning classes, and predicting.

Steps involved in decision stump

Step 1. Selecting the feature
Step 2. Choosing the threshold
Step 3. Assigning class labels
Step 4. Making predictions
Step 5. Training and evaluation

2.3 M5P

M5P is a machine learning algorithm used for regression tasks. It is an extension of the decision tree-based model called M5, which Ross Quinlan developed. The M5 algorithm combines decision trees and linear regression to create more accurate and flexible regression models. M5P, specifically, stands for M5 Prime. It enhances the original M5 algorithm to improve its predictive performance. M5P uses a tree-based model to divide the data into subsets based on feature values recursively and then fits linear regression models to each of these subsets. The result is a piecewise linear regression model, where different linear regressions are used for other regions of the input feature space.

Steps involved in the M5P

Step 1. Building the initial decision tree (M5 model): Recursive Binary Splitting and Pruning (optional)
Step 2. Linear Regression Model: Leaf Regression Models and Model Parameters
Step 3. Piecewise Linear Regression: Piecewise Prediction
Step 4. Model Evaluation: Training and Testing.

2.4 Random Forest

Random Forest is a popular machine learning ensemble method for classification and regression tasks. It is an extension of decision trees and is known for its high accuracy, robustness, and ability to handle complex datasets. Random Forest is widely used in various domains, including data science, machine learning, and pattern recognition. The main idea behind Random Forest is to create an ensemble (a collection) of decision trees and combine their predictions to make more accurate and stable predictions. The following steps describe what Random Forest works like.

- Bootstrap Aggregating (Bagging)
- Decision Tree Construction
- Voting for Classification, Averaging for Regression

The key advantages of Random Forest are:

- Reduced overfitting
- Robustness
- Feature Importance

Steps involved in Random Forest

Random Forest is an ensemble learning method combining multiple decision trees to make more accurate and robust predictions for classification and regression tasks. The steps involved in building a Random Forest are as follows:

Step 1. Data Bootstrapping
Step 2. Random Feature Subset Selection
Step 3. Decision Tree Construction
Step 4. Ensemble of Decision Trees
Step 5. Out-of-Bag (OOB) Evaluation
Step 6. Hyperparameter Tuning (optional)

2.5 Random Tree

In machine learning, a Random Tree is a specific type of decision tree variant that introduces randomness during construction. Random Trees are similar to traditional decision trees but differ in how they select the splitting features and thresholds at each node. The primary goal of introducing randomness is to create a more diverse set of decision trees, which can help reduce overfitting and improve the model's generalization performance. Random Trees are commonly used as building blocks in ensemble methods like Random Forests. The critical characteristics of Random Trees are as follows:

- Random Feature Subset
- Random Threshold Selection
- No Pruning
- Ensemble Methods

Steps involved in Random Tree

Step 1. Data Bootstrapping:

Step 2. Random Subset Selection for Features:
Step 3. Decision Tree Construction:
Step 4. Voting (Classification) or Averaging (Regression):

2.6 REP Tree

REP (Repeated Incremental Pruning to Produce Error Reduction) Tree is a machine learning algorithm for classification and regression tasks. A decision tree-based algorithm constructs a decision tree using a combination of incremental pruning and error-reduction techniques. The key steps involved in building a REP Tree are as follows:

- ❖ Recursive Binary Splitting
- ❖ Pruning
- ❖ Repeated Pruning and Error Reduction

Steps involved in REP Tree

REP Tree (Repeated Incremental Pruning to Produce an Error Reduction Tree) is a machine learning algorithm for classification and regression tasks. It is an extension of decision trees that incorporates pruning to reduce overfitting and improve the model's generalization performance. Below are the steps involved in building a REP Tree.

Step 1. Recursive Binary Splitting
Step 2. Pruning
Step 3. Repeated Pruning and Error Reduction
Step 4. Model Evaluation

2.8 Accuracy Metrics

The predictive model's error rate can be evaluated by applying several accuracy metrics in machine learning and statistics. The basic concept of accuracy evaluation in regression analysis is comparing the original target with the predicted one and using metrics like R-squared, MAE, MSE, and RMSE to explain the errors and predictive ability of the model [14]. The R-squared, MSE, MAE, and RMSE are metrics used to evaluate the prediction error rates and model performance in analysis and predictions [15] and [16].

R-squared (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The values from 0 to 1 are interpreted as percentages. The higher the value is, the better the model is.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2} \quad \ldots (2)$$

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad \ldots (3)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad \ldots (4)$$

Relative Absolute Error (RAE) is a metric used in statistics and data analysis to measure the accuracy of a forecasting or predictive model's predictions. It is particularly useful when dealing with numerical data, such as in regression analysis or time series forecasting.

$$\mathrm{RAE} = \frac{\Sigma|y_i - \hat{y}_i|}{\Sigma|y_i - \bar{y}|} \quad \ldots (5)$$

Root Relative Squared Error (RRSE) is another metric used in statistics and data analysis to evaluate the accuracy of predictive models, especially in the context of regression analysis or time series forecasting.

$$\mathrm{RRSE} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}} \quad \ldots (6)$$

Equation 3 to 7 are used to find the model accuracy, which is used to find the model performance and error. Where Y_i represents the individual observed (actual) values, $\hat{Y}_i$ represents the corresponding individual predicted values, $\bar{Y}$ represents the mean (average) of the observed values and Σ represents the summation symbol, indicating that you should sum the absolute differences for all data points.

Numerical Illustrations

The diabetes_prediction_dataset.csv
file contains medical and demographic data of patients along with their diabetes status, whether positive or negative. It consists of various features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The Dataset can be utilized to construct machine learning models that can predict the likelihood of diabetes in patients based on their medical history and demographic details [17].

Table 1. Diabetes prediction sample dataset

gender	age	hypertension	Heart disease	Smoking history	bmi	HbA1c level	Blood Glucose level	diabetes

Female	80	0	1	Never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	Never	27.32	5.7	158	0
Female	36	0	0	Current	23.45	5	155	0
Male	76	1	1	Current	20.14	4.8	155	0
Female	44	0	0	never	19.31	6.5	200	1
Female	79	0	0	No Info	23.86	5.7	85	0
Male	42	0	0	never	33.64	4.8	145	0
Female	32	0	0	never	27.32	5	100	0
Female	53	0	0	never	27.32	6.1	85	0
Female	53	0	0	former	27.32	7	159	1
Female	45	1	0	never	23.05	4.8	130	0
Male	50	0	0	former	37.16	9	159	1
Male	30	0	0	No Info	27.32	6.6	140	0
Female	19	0	0	never	23.35	3.5	155	0
Male	46	0	0	No Info	24.41	5	140	0
Female	67	0	0	never	63.48	8.8	155	1

Table 2: Machine Learning Models with Correlation coefficient

ML Approaches	**Correlation coefficient**
Linear Regression	0.5906
Decision Stump	0.6605
M5P	0.8321
Random Forest	0.8211
Random Tree	0.6946
REP Tree	0.8296

Table 3: Machine Learning Models with MAE and RMSE

ML Approaches	**MAE**	**RMSE**
Linear Regression	0.1538	0.2251
Decision Stump	0.0877	0.2094
M5P	0.0485	0.1547
Random Forest	0.0475	0.1596
Random Tree	0.0485	0.2193
REP Tree	0.0460	0.1558

Table 4: Machine Learning Models with RAE (%) and RRSE (%)

ML Approaches	**RAE (%)**	**RRSE (%)**
Linear Regression	98.8652	80.6968
Decision Stump	56.3720	75.0810
M5P	31.1532	55.4569
Random Forest	30.5576	57.2316

Random Tree	31.1553	78.6220
REP Tree	29.5894	55.8757

Table 5: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Time taken
Linear Regression	2.7600
Decision Stump	0.8300
M5P	15.5000
Random Forest	81.3900
Random Tree	0.9500
REP Tree	1.8100

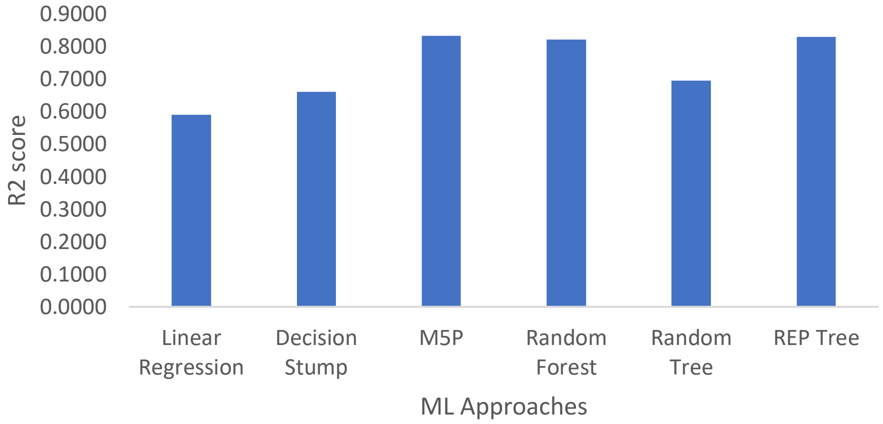

Fig. 1. R2 Score for Machine Learning Approaches

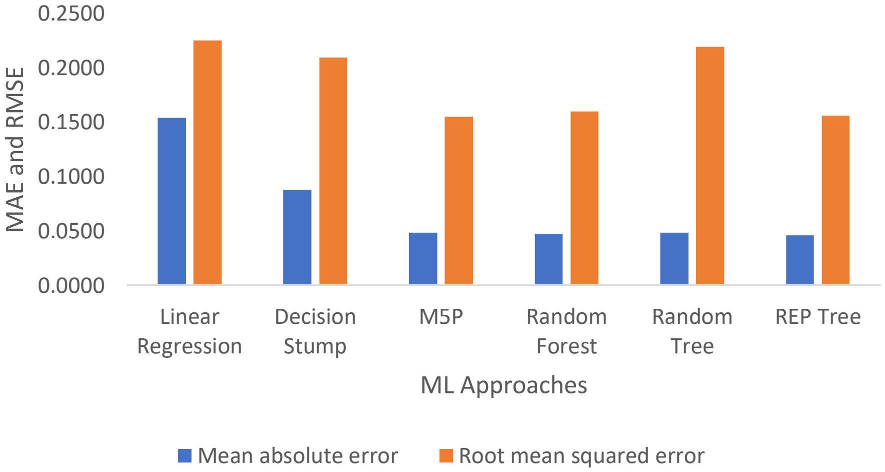

Fig. 2. Machine Learning Models with MAE and RMSE

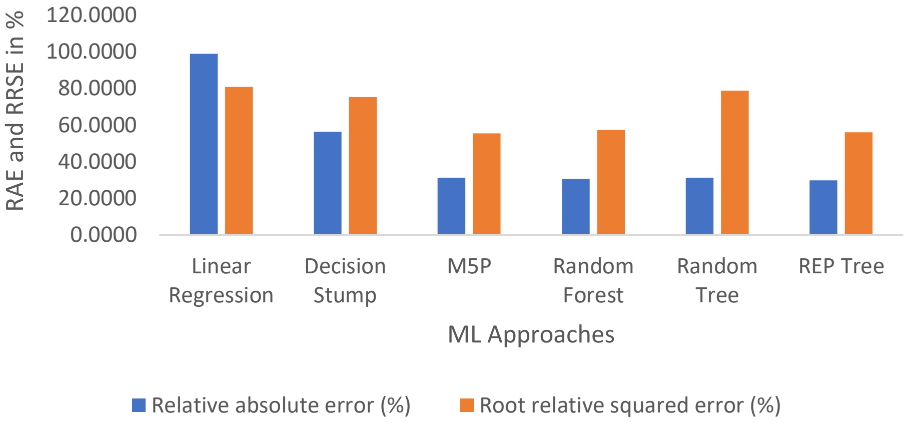

Fig. 3. Machine Learning Models with RAE (%) and RRSE (%)

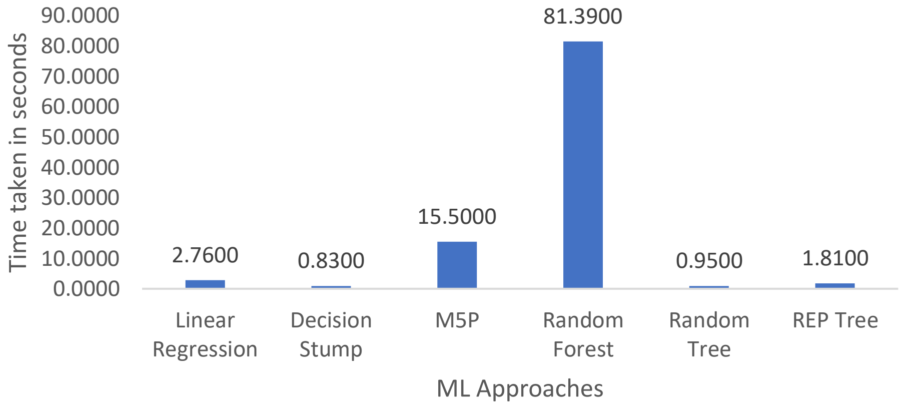

Fig. 4. Machine Learning Models and its Time Taken to Build the Model (Seconds)

3. Results and Discussion

In Table 1, we elaborate on nine parameters, encompassing various data categories, such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, Blood Glucose level, and diabetes. Our dataset indicates the utilization of six additional machine learning approaches specifically, linear regression, decision stump, M5P, random forest, random tree, and REP tree to unveil concealed patterns and identify the most influential parameters for future predictions. We present the relevant findings and numerical representations in Tables 1 to 5 and Figures 1 to 4.

These outcomes are based on Equation 2, Table 2, and Figure 1, which facilitate the computation of the R2 score or correlation coefficient across the nine parameters. The numerical results suggest noteworthy variations between these parameters. Notably, when analyzing the death event, all six machine learning approaches demonstrate a strong positive correlation of approximately 0.5.

We utilize the Mean Absolute Error (MAE) calculated using Equation 3 to assess model errors in the context of six machine learning algorithms. It is noteworthy that all six machine learning approaches exhibit minimal error performance, approaching nearly zero. Similarly, we employ the Root Mean Square Error (RMSE) described in Equation 4 to quantify the disparities between predicted and actual values, with all six ML approaches yielding near-zero error performance. Detailed numerical representations can be found in Table 3 and Figure 2.

The Relative Absolute Error (RAE), as defined by Equation 5, serves to gauge accuracy by assessing the divergence between predicted and actual values in percentage terms. In our research, involving six ML classification algorithms, Linear Regression presents the highest error rate, while the remaining five ML approaches excel in minimizing errors. This pattern is consistent when evaluating the Relative Root Square Error (RRSE), as supported by the data in Table 4 and Figure 3.

Furthermore, we consider the time taken a crucial aspect of machine learning approaches. According to Table 5 and Figure 4, M5P and Random Forests require the most time to resolve the problem, while Decision Stump, Random Tree, and REP Tree exhibit minimal time requirements for model construction. Notably, Linear Regression also boasts an efficient time utilization, aligning with the visual representations provided.

Diabetes prediction is based on various parameters for using machine learning approaches. In this case the ML approaches return various accuracy parameters. Based on six ML approaches namely linear regression, decision stump, M5P, random forest, random tree, and REP tree, LR, DS, and RT return normal positive correlation. Remaining approaches namely RF and REP Tree return very strong positive correlation with best accuracy performance. In this research clearly states that, all the parameters influenced to the prediction of diabetes.

4. Conclusion and Future Research

In conclusion, this study addresses various aspects and constraints related to our model, including gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, Blood Glucose level, and diabetes data-specific considerations. We also acknowledge factors that may influence potential underperformance and computational limitations in model development. We recommend potential enhancements and future steps, such as exploring additional data sources, investigating superior algorithms and hyperparameters, and fine-tuning the model to elevate its performance. Our research in diabetes is characterized by a dynamic collaboration among scientists, healthcare professionals, and patients, with an unwavering commitment to enhancing the well-being of individuals living with diabetes and reducing its societal impact.

Reference

1. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, pp.104-116.
2. Kumari, S. and Singh, A., 2013, January. A data mining approach for the diagnosis of diabetes mellitus. In 2013 7th International Conference on Intelligent Systems and Control (ISCO) (pp. 373-375). IEEE.
3. Georga, E., Protopappas, V., Guillen, A., Fico, G., Ardigo, D., Arredondo, M.T., Exarchos, T.P., Polyzos, D. and Fotiadis, D.I., 2009, September. Data mining for blood glucose prediction and knowledge discovery in diabetic patients: The METABO diabetes modeling and management system. In 2009 annual international conference of the IEEE engineering in medicine and biology society (pp. 5633-5636). IEEE.
4. Rajesh, K. and Sangeetha, V., 2012. Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT), 2(3).
5. Georga, E.I., Protopappas, V.C., Mougiakakou, S.G. and Fotiadis, D.I., 2013, November. Short-term vs. long-term analysis of diabetes data: Application of machine learning and data mining techniques. In 13th IEEE International Conference on BioInformatics and BioEngineering (pp. 1-4). IEEE.
6. Rajesh, P. and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using the Weka tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243.
7. Shivakumar, B.L. and Alby, S., 2014, March. A survey on data-mining technologies for prediction and diagnosis of diabetes. In 2014 International Conference on Intelligent Computing Applications (pp. 167-173). IEEE.
8. Sanakal, R. and Jayakumari, T., 2014. Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. International Journal of Computer Trends and Technology, 11(2), pp.94-98.
9. Yıldırım, E.G., Karahoca, A. and Uçar, T., 2011. Dosage planning for diabetes patients using data mining methods. Procedia Computer Science, 3, pp.1374-1380.
10. Rajesh, P., Karthikeyan, M. and Arulpavai, R., 2019, December. Data mining approaches to predict the factors that affect the groundwater level using a stochastic model. In AIP Conference Proceedings (Vol. 2177, No. 1). AIP Publishing.
11. Rajesh, P. and Karthikeyan, M., 2019. Data mining approaches to predict the factors that affect agriculture growth using stochastic models. International Journal of Computer Sciences and Engineering, 7(4), pp.18-23.
12. Rajesh, P., Karthikeyan, M., Santhosh Kumar, B. and Mohamed Parvees, M.Y., 2019. Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience, 16(4), pp.1472-1477.
13. Kohavi, R., & Sahami, M. (1996). Error-based pruning of decision trees. In International Conference on Machine Learning (pp. 278-286).

14. Akusok, A. (2020). What is Mean Absolute Error (MAE)? Retrieved from https://machinelearningmastery.com/mean-absolute-error-mae-for-machine-learning/

15. S. M. Hosseini, S. M. Hosseini, and M. R. Mehrabian, “Root mean square error (RMSE): A comprehensive review,” International Journal of Applied Mathematics and Statistics, vol. 59, no. 1, pp. 42–49, 2019.

16. Chi, W. (2020). Relative Absolute Error (RAE) – Definition and Examples. Medium. https://medium.com/@wchi/relative-absolute-error-rae-definition-and-examples-e37a24c1b566

17. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset