A Study on the Effectiveness of Classifiers in Diagnosing Diabetes Using Machine and Deep Learning Models

Dr. R. Raja[1], Dr. P. Rajesh[2]

[1]*Assistant Professor, Dept. of Computer Science, Thiru Kolanjiappar Govt. Arts College, (Deputed from Dept. of Computer and Information Science, Annamalai University, Annamalainagar-608 002) Tamil Nadu, India, rajamanira2000@gmail.com*
[2]*Assistant Professor, PG Department of Computer Science, Government Arts College, (Deputed from Dept. of Computer and Information Science, Annamalai University, Annamalainagar-608 002) Tamil Nadu, India, rajeshdatamining@gmail.com*

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

ABSTRACT
Machine learning and data mining have a crucial role in assisting in the early detection of diabetes, potentially resulting in improved health outcomes and more effective management for individuals at risk. This process is frequently termed diabetes prediction or diabetes risk assessment. This paper considers diabetes another-related dataset for applying data mining techniques to find suitable variables for future predictions. Machine learning algorithms can be harnessed in Industrial Internet of Things (IIoT) applications to unlock the advantages of cost reduction, enhanced efficiency, and improved performance. In the modern era, we've all witnessed the benefits of machine learning techniques, from streaming movie services suggesting titles based on viewing habits to identifying fraudulent activity through customer spending patterns. These algorithms excel at handling vast and intricate datasets, uncovering intriguing patterns and trends, including anomalies. This paper considers diabetes another-related dataset data data like age, gender, family diabetes, highbp, physically active, bmi, smoking, alcohol, sleep, sound sleep, regular medicine, junkfood, stress, bplevel, pregancies, pdiabetes, uriationfreq, diabetic. The machine learning approaches which is used to analysis and predict the dataset using Logistic, Multilayer Perceptron, SMO, Decision Stump, Hoeffding Tree, J48, and LMT. Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

1. Introduction and Background

Utilizing machine learning and data mining methods in diabetes research can result in invaluable discoveries, enhanced diagnostic precision, and superior patient care. Within this field, researchers are persistently seeking novel approaches that harness data-driven methods to achieve a deeper comprehension and more effective management of diabetes.

Machine learning algorithms employ parameters derived from training data, which is a subset representing the broader dataset. As the training data expands to better mirror the real world, the algorithm can produce increasingly precise results. The techniques encompass pattern tracking, classification, association, outlier detection, clustering, regression, and prediction. Identifying patterns is straightforward, often facilitated by abrupt data changes. We've gathered and organized data into distinct sections for in-depth category-based analysis.

The objective of this study is to conduct a systematic review of the utilization of machine learning, data mining techniques, and tools in the realm of diabetes research, focusing on a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management, with the initial category appearing to be the most prevalent. A diverse array of machine learning algorithms were applied. In general, 85% of these employed supervised learning approaches, while 15% utilized unsupervised methods, particularly association rules. Support Vector Machines (SVM) emerged as the most successful and widely used algorithm. Clinical datasets constituted the primary data type. The titles of the selected articles underscore the value of extracting knowledge to drive deeper understanding and further exploration in diabetes research [1].

The realm of medical diagnosis, a reliable prediction methodology for diabetes is indispensable. Data mining, which involves analyzing data from multiple angles and summarizing it into valuable information, plays a pivotal role. The primary objective of data mining is to uncover fresh patterns and provide meaningful insights to users. This paper endeavors to mine relationships within diabetes data for effective classification. Various data mining methods and techniques will be explored to identify the most suitable approaches for efficiently classifying diabetes datasets and unearthing valuable patterns [2].

Chronic diabetes care entails extensive data on self-management and clinical aspects of the condition. This paper proposes two distinct approaches. Firstly, it presents a predictive model for short-term glucose homeostasis that relies on machine learning to prevent hypoglycemic events and extended hyperglycemia on a daily basis. Secondly, data mining methods are suggested to elucidate and predict long-term glucose control and the incidence of diabetic complications [3].

The World Health Organization's recent report underscores the increasing prevalence of diabetes worldwide. Early identification is a significant challenge. Data mining has played a crucial role in diabetes research by extracting hidden knowledge from extensive diabetes-related datasets. Various data mining techniques contribute to diabetes research, ultimately enhancing healthcare quality for diabetic patients. This paper offers a survey of commonly applied data mining methods for diabetes data analysis and disease prediction [4].

Clinical decision-making necessitates available information to guide physicians. Currently, data mining methods are applied in medical research to analyze extensive medical data. This study seeks to utilize data mining to analyze a database of diabetes cases and diagnose the disease. It involves implementing Fuzzy C-Means (FCM) and Support Vector Machine (SVM) and testing them on a dataset related to diabetes diagnosis. The dataset comprises nine input attributes related to clinical diabetes diagnosis and one output attribute indicating whether the patient has diabetes. The entire dataset comprises 768 cases [5].

Data mining serves as a valuable tool for exploring large pre-existing databases to uncover previously unknown valuable information. In this paper, a weather dataset with attributes such as Outlook, Temperature, Humidity, Windy, and Boolean Play Golf class variables is used for training. Seven classification algorithms, including J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP), and Random Forest (RF), are employed to measure accuracy. Among these, the Random Tree algorithm outperforms the others with an accuracy of 85.714% [6].

High blood sugar, known as diabetes mellitus, can result from insufficient insulin production or improper cell response to insulin. This study aims to develop a data mining model to predict suitable dosage plans for diabetes patients. The study utilizes medical records from 89 patients, comprising 318 diabetes assays. The Adaptive Neuro-Fuzzy Inference System (ANFIS) and Rough Set methods are employed for dosage planning. According to the results, ANFIS is the more successful and reliable method for diabetes drug planning compared to Rough Set [7].

Data mining techniques (DMT) provide powerful tools for extracting knowledge from data, aiding in decision-making. This paper explores the use of DMT in diabetes self-management (DSM). The study conducts a systematic mapping analysis to review primary studies related to DMT in DSM. The analysis covers years and sources of DSM publications, the most studied type of diabetes, the most frequently used DM tasks and techniques, and the considered functionalities. Out of 57 selected papers published between 2000 and April 2017, prediction is the most frequently used DM task, and Neural Networks are the most commonly applied technique. Additionally, Type 1 diabetes (T1DM) receives significant attention in these studies [8]. Data mining offers a promising avenue for early prediction of diabetes, a chronic disease that affects various organs. This paper explores early diabetes prediction using various data mining techniques. The dataset comprises 768 instances from the PIMA Indian Dataset. The analysis reveals that the Modified J48 Classifier provides the highest accuracy compared to other techniques [9].

Data mining can efficiently harness stochastic sensing for predictive assessments. This paper effectively assesses groundwater levels, rainfall, population, food grains, and enterprises through stochastic modeling and data mining. It introduces a novel data assimilation analysis to predict groundwater levels effectively. Experimental results demonstrate the robustness of this approach [10] and [11]. In another dataset, attributes represent topics, questions, data values, low confidence limits, and high confidence limits. This data is used for training and testing with five classification algorithms. The study evaluates and compares the accuracy of five different decision tree algorithms, with the M5P decision tree approach outperforming the others [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 Logistic Regression

Logistic Regression is a supervised learning algorithm often used for binary classification. It predicts the probability that a given input belongs to a specific class, using the logistic function to map predictions between 0 and 1 (Hosmer et al., 2013). The process is as follows:

Step 1. Data Preparation: Data is pre-processed, and features are normalized if needed.

Step 2. Model Initialization: Initialize weights and bias parameters.

Step 3. Hypothesis Calculation: Apply a linear combination of the input features and weights.

Step 4. Cost Function: Use the binary cross-entropy loss to measure error in predictions.

Step 5. Gradient Descent Optimization: Adjust weights iteratively by calculating the gradient

of the cost function concerning each parameter to minimize error.

Step 6. Prediction: After training, use the model to classify new data points based on learned weights.

2.2 Multilayer Perception

A Multilayer Perceptron (MLP) is an artificial neural network consisting of multiple layers of interconnected nodes or neurons. It's a fundamental architecture in deep learning and is used for various tasks, including classification, regression, and more complex tasks like image recognition and natural language processing. The architecture of an MLP typically includes three types of layers:

i. Input Layer
ii. Hidden Layers
iii. Output Layer

2.3 SMO

SMO stands for "Sequential Minimal Optimization," an algorithm used for training support vector machines (SVMs), machine learning models commonly used for classification and regression tasks. The SMO algorithm is particularly well-suited for solving the quadratic programming optimization problem that arises during the training of SVMs.

Step 1. Initialization

Step 2. Selection of Two Lagrange Multipliers

Step 3. Optimize the Pair of Lagrange Multipliers

Step 4. Update the Model

Step 5. Convergence Checking

Step 6. Repeat

2.4 Decision Stump

A Decision Stump is a simple machine learning model that serves as a weak learner, often used in ensemble learning methods like boosting. It's a basic model that makes decisions based on a single feature (input) and a threshold value. Despite its simplicity, when combined with other decision stumps or more complex models, decision stumps can contribute to building stronger predictive models. Here's how a Decision Stump works:

Step 1. Input Feature

Step 2. Threshold

Step 3. Prediction

Step 4. Decision Rule

2.5 Hoeffding Tree

A Hoeffding Tree, also known as VFDT (Very Fast Decision Tree) or Incremental Decision Tree, is a machine learning algorithm designed for online, incremental learning on streaming data. It's beneficial when you have large volumes of data that are continuously arriving and you want to update your model in real-time without retraining the entire dataset. Here's a simplified overview of how the Hoeffding Tree algorithm works:

Step 1. Initialization

Step 2. Data Arrival

Step 3. Splitting Nodes

Step 4. Leaf Node Prediction

Step 5. Adaptation

2.6 J48

J48, also known as C4.5, is a popular decision tree algorithm used for classification tasks in machine learning and data mining. It was developed by Ross Quinlan and is an extension of the earlier ID3 (Iterative Dichotomiser 3) algorithm. J48 is widely used due to its effectiveness, ease of use, and ability to handle both categorical and numerical attributes. Here are the key features and steps of the J48 algorithm:

Step 1. Attribute Selection

Step 2. Splitting Nodes

Step 3. Recursion

Step 4. Pruning

Step 5. Handling Missing Values

Step 6. Post-Pruning

Step 7. Leaf Node Prediction

2.7 LMT

LMT (Logistic Model Trees) is a machine learning algorithm that combines decision trees with logistic regression to create a hybrid model for classification tasks. It aims to harness the strengths of both decision trees and logistic regression, mitigating their individual weaknesses. LMT was introduced as an alternative to traditional decision trees and has shown promise in improving predictive performance and interpretability. Here's how the LMT algorithm works:

Step 1. Decision Tree Generation

Step 2. Leaf Node Transformation

Step 3. Predictions

2.8 Kappa statistic

The Kappa statistic, also called Cohen's Kappa or simply Kappa, is a statistical metric utilized to assess the level of agreement between two or more raters or classifiers when assigning categorical ratings or labels to items. It goes beyond considering agreement by chance alone. The Kappa statistic is represented on a scale from -1 to 1. A Kappa value of -1 signifies perfect disagreement between the raters or classifiers. A Kappa value of 0 indicates agreement that is no better than chance. A Kappa value of 1 implies perfect agreement between the raters or classifiers. The calculation of Kappa employs the formula:

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e}$$

$$P_o = \frac{\text{Number of items with agreement}}{\text{Total number of items}}$$

$$P_e = \sum \frac{\text{Total count in row} \times \text{Total count in column}}{\text{Total number of items}}$$

Where, P_o denotes the observed agreement, i.e., the proportion of items on which raters or classifiers agree. P_e

represents the expected agreement, i.e., the agreement expected by chance.

2.9 Mean Absolute Error

Mean Absolute Error (MAE) is a metric used to measure the average absolute difference between predicted and actual (true) values in a regression problem. It is commonly used to assess the accuracy of a regression model's predictions [14]. The formula to calculate Mean Absolute Error (MAE) is as follows:

$$MAE = \Sigma |(\text{Actual Value} - \text{Predicted Value})| / n \quad \ldots (2)$$

Where:

- Σ represents the summation symbol, which sums up the values for all data points.
- | | denotes the absolute value, ensuring the differences are positive.

In this formula:

- Actual Value: Refers to the true value of the target variable (ground truth) for a specific data point.
- Predicted Value: Refers to the value predicted by the regression model for the same data point.
- n: Represents the total number of data points in the dataset.

2.10 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a commonly used metric to assess the accuracy of a regression model's predictions. It measures the average magnitude of the errors between the predicted and actual (true) values, considering both the direction and magnitude of the errors. The formula to calculate Root Mean Squared Error (RMSE) is as follows [15]:

$$RMSE = \sqrt{(\Sigma (\text{Actual Value} - \text{Predicted Value})^2 / n)} \quad \ldots (3)$$

Where:

- Σ represents the summation symbol, which sums up the values for all data points.
- (Actual Value - Predicted Value) ² denotes the squared difference between the actual and predicted values for each data point.
- n is the total number of data points in the dataset.

2.11 Relative Absolute Error (RAE)

Relative Absolute Error (RAE), also known as Mean Absolute Percentage Error (MAPE), is a metric used to evaluate the accuracy of predictions in regression tasks. It measures the average percentage difference between the absolute and actual (valid) values, providing a relative measure of the prediction errors [16]. The formula to calculate Relative Absolute Error (RAE) is as follows:

$$RAE = (\Sigma |\text{Actual Value} - \text{Predicted Value}| / \Sigma |\text{Actual Value}|) * (100 / n) \quad \ldots (4)$$

Where:

- Σ represents the summation symbol, which sums up the values for all data points.
- | | denotes the absolute value, ensuring the differences are positive.
- n is the total number of data points in the dataset.

2.12 Root Relative Squared Error (RRSE)

"Root Relative Squared Error" is not a standard or widely recognized metric in statistics or machine learning. It appears to be a combination of the terms "Root Mean Squared Error (RMSE)" and "Relative Absolute Error (RAE)." It's possible that the time was created or used in a specific context or literature, but it is not a commonly used or established metric. For clarity, let's briefly define the two individual metrics mentioned:

Root Mean Squared Error (RMSE): As explained earlier, RMSE is a commonly used metric to evaluate the accuracy of regression models. It measures the average magnitude of the errors between the predicted and actual values, considering both the direction and extent of the errors. The formula to calculate RMSE is:

RMSE = √(Σ (Actual Value - Predicted Value)² / n)

❖ Relative Absolute Error (RAE): Also known as Mean Absolute Percentage Error (MAPE), RAE measures the average percentage difference between the absolute errors and the actual (true) values, providing a relative measure of the prediction errors. The formula to calculate RAE is:

RAE = (Σ |Actual Value - Predicted Value| / Σ |Actual Value|) * (100 / n)

As there is no established metric called "Root Relative Squared Error," it's crucial to use standard evaluation metrics such as RMSE, RAE (MAPE), or others that are well-known and have clear interpretations in the context of your specific problem.

3. Numerical Illustrations

The corresponding dataset was collected from the open souse Kaggle data repository. The diabetes another dataset includes 18 parameters which have different categories of data like age, gender, family diabetes, highbp, physically active, bmi, smoking, alcohol, sleep, sound sleep, regular medicine, junkfood, stress, bplevel, pregancies, pdiabetes, uriationfreq, diabetic [17]. A detailed description of the parameters is mentioned in the following Table 1.

Table 1. Diabetes another sample dataset

Age	Gender	Family Diabetes	highBP	Physically Active	BMI	Smoking	Alcohol	Sleep	Sound Sleep	Regular Medicine	JunkFood	Stress	BP Level	Pregancies	P diabetes	Uriation Freq	Diabetic
50-59	Male	No	yes	one hr or more	39	no	No	8	6	no	occasionally	sometimes	high	0	0	not much	no
50-59	Male	No	yes	less than half an hr	28	no	No	8	6	yes	very often	sometimes	normal	0	0	not much	no
40-49	Male	No	no	one hr or more	24	no	No	6	6	no	occasionally	sometimes	normal	0	0	not much	no
50-59	Male	No	no	one hr or more	23	no	No	8	6	no	occasionally	sometimes	normal	0	0	not much	no
40-49	Male	No	no	less than half an hr	27	no	No	8	8	no	occasionally	sometimes	normal	0	0	not much	no
40-49	Male	No	yes	none	21	no	Yes	10	10	no	occasionally	sometimes	high	0	0	not much	yes
less than 40	Male	No	no	one hr or more	24	no	No	8	8	no	occasionally	sometimes	normal	0	0	not much	no
less than 40	Male	No	no	less than half an hr	20	no	No	7	7	yes	occasionally	sometimes	low	0	0	not much	no
40-49	Male	yes	no	one hr or more	23	no	No	7	7	no	occasionally	sometimes	normal	0	0	not much	no
less than 40	Male	No	no	more than half an hr	20	no	No	8	8	o	occasionally	sometimes	normal	0	0	not much	no
less than 40	Male	No	no	none	20	no	No	7	7	no	occasionally	not at all	normal	0	0	not much	no
40-49	Male	No	no	less than half an hr	26	yes	No	8	7	no	occasionally	sometimes	normal	0	0	not much	no
less than 40	Female	No	no	less than half an	21	no	No	6	6	no	occasionally	sometimes	normal	1	0	not much	no

				hr													
less than 40	Female	No	no	one hr or more	22	no	No	8	7	no	occasionally	very often	normal	2	0	not much	no
less than 40	Male	No	no	one hr or more	15	no	No	7	7	no	occasionally	very often	normal	0	0	not much	no
40-49	Female	yes	no	none	34	no	No	6	6	no	occasionally	sometimes	normal	3	0	not much	no

Table 2: Machine Learning Models with Correctly Classified Instances and Incorrectly Classified Instances

ML Approaches	Correctly Classified Instances	Incorrectly Classified Instances
Logistic	845.0000	106.0000
Multilayer Perceptron	899.0000	52.0000
SMO	836.0000	115.0000
Decision Stump	779.0000	172.0000
Hoeffding Tree	798.0000	153.0000
J48	897.0000	54.0000
LMT	910.0000	41.0000

Table 3: Machine Learning Models with Correctly Classified Instances (%) and Incorrectly Classified Instances (%)

ML Approaches	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)
Logistic	88.8538	11.1462
Multilayer Perceptron	94.5321	5.4679
SMO	87.9075	12.0925
Decision Stump	81.9138	18.0862
Hoeffding Tree	83.9117	16.0883
J48	94.3218	5.6782
LMT	95.6887	4.3113

Table 4: Machine Learning Models with Kappa statistic

ML Approaches	Kappa statistic
Logistic	0.7184
Multilayer Perceptron	0.8612
SMO	0.6949
Decision Stump	0.5857
Hoeffding Tree	0.6147
J48	0.8555
LMT	0.8917

Table 5: Machine Learning Models with Mean Absolute and Root Mean Squared Error

ML Approaches	MAE	RMSE
Logistic	0.1187	0.2459
Multilayer Perceptron	0.0449	0.1782
SMO	0.2493	0.3181
Decision Stump	0.1744	0.2956
Hoeffding Tree	0.1179	0.2959
J48	0.0592	0.1855
LMT	0.0386	0.1529

Table 6: Machine Learning Models with Relative Absolute Error (%) and Root Relative Squared Error (%)

ML Approaches	RAE (%)	RRSE (%)
Logistic	43.9171	66.9569
Multilayer Perceptron	16.6129	48.5225
SMO	92.2582	86.6292
Decision Stump	64.5154	80.4995
Hoeffding Tree	43.6146	80.5737
J48	21.8935	50.5318
LMT	14.2786	41.6526

Table 7: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Time taken (seconds)
Logistic	0.3700
Multilayer Perceptron	8.8000
SMO	0.4700
Decision Stump	0.0100
Hoeffding Tree	0.0900
J48	0.1200
LMT	1.8300

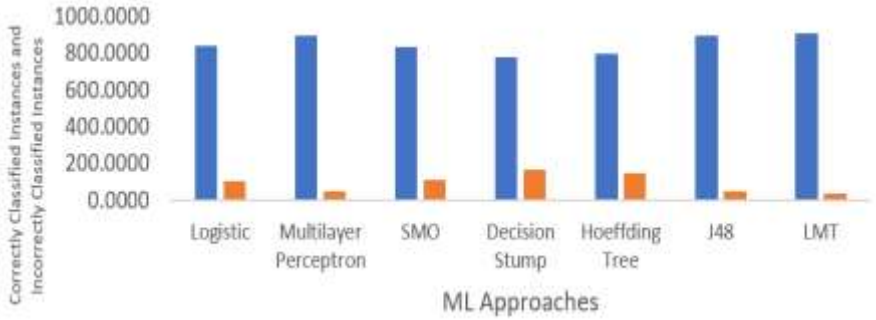

Fig. 1. Machine Learning Models with Correctly Classified Instances and Incorrectly Classified Instances

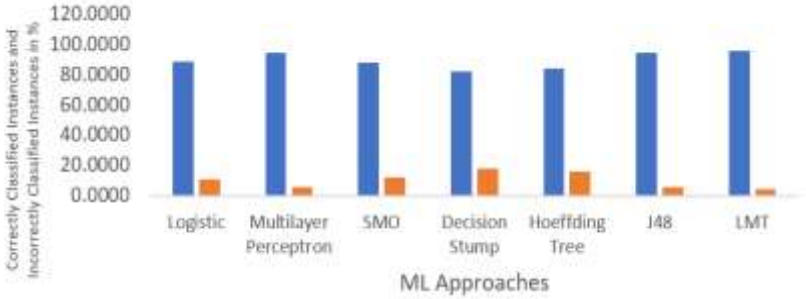

Fig. 2. Machine Learning Models with Correctly Classified Instances (%) and Incorrectly Classified Instances (%)

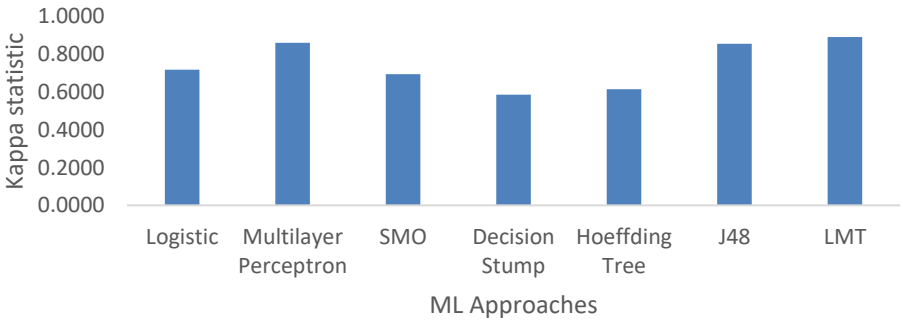

Fig. 3. Machine Learning Models with Kappa statistic

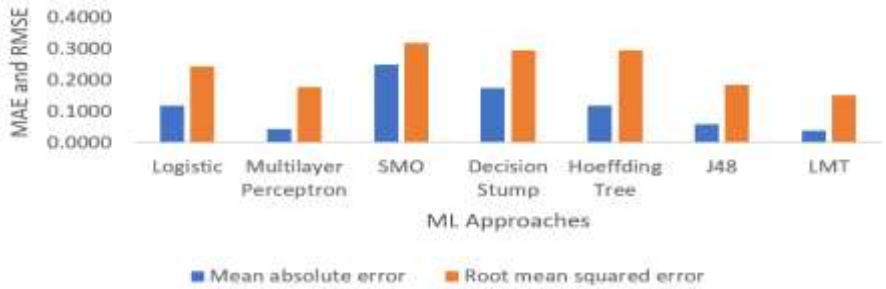

Fig. 4. Machine Learning Models with MAE and RMSE

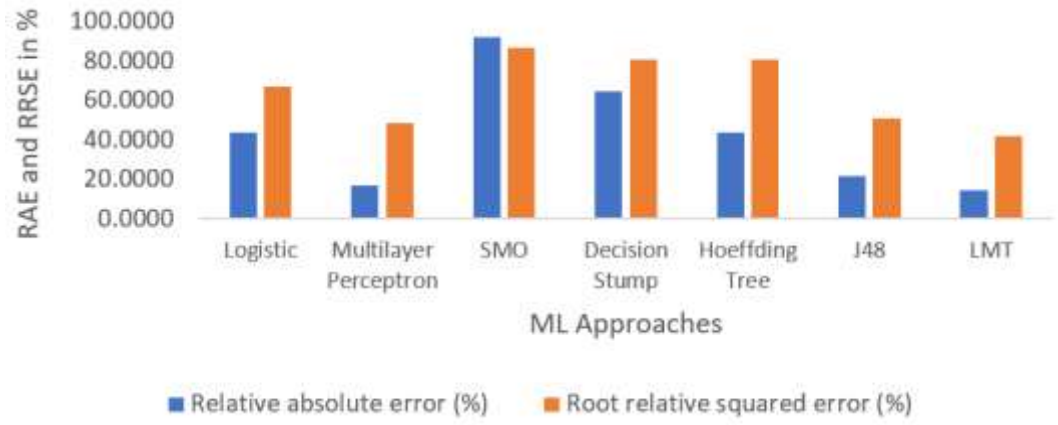

Fig. 5. Machine Learning Models with RAE (%) and RRSE (%)

Fig. 6. Machine Learning Models and its Time Taken to Build the Model (Seconds)

Table 8: ML Approaches performance

ML Approachs	Logistc	Multilayr Perceptrn	SMO	Decision Stump	Hoeffding Tree	J48	LMT
TP Rate	0.5690	0.6143	0.5613	0.5450	0.5450	0.6117	0.6273
FP Rate	0.0987	0.0537	0.1063	0.1227	0.1220	0.0560	0.0403
Precision	0.5783	0.6290	0.5700	0.5193	0.5350	0.6117	0.6353
Recall	0.5690	0.6143	0.5613	0.5450	0.5450	0.6117	0.6273
F-Measure	0.5733	0.6210	0.5653	0.5277	0.5393	0.6190	0.6310
MCC	0.4800	0.5760	0.4643	0.3967	0.4120	0.5723	0.5957
ROC Area	0.8943	0.7770	0.7250	0.5993	0.6070	0.7860	0.7307
PRC Area	0.6003	0.6327	0.5230	0.4803	0.5810	0.6313	0.6477

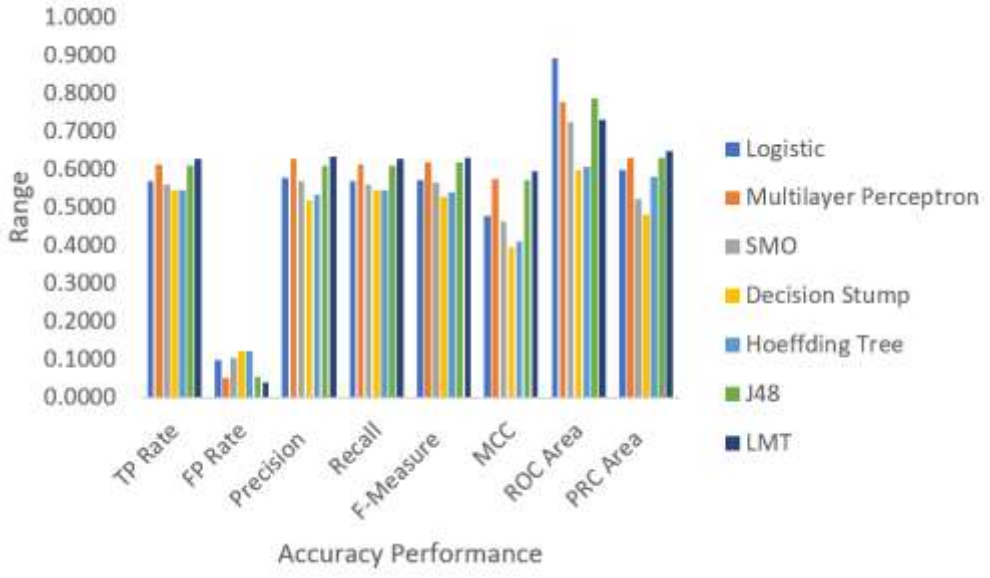

Fig. 7: ML Approaches performance

4. Results and Discussion

The findings of our study. Table 1 provides a comprehensive overview of 18 parameters, encompassing various data categories, such as age, gender, family history of diabetes, high blood pressure, physical activity, body mass index (BMI), smoking, alcohol consumption, sleep patterns, sound sleep, regular medication, junk food intake, stress levels, blood pressure readings, pregnancies, pre-existing diabetes, urination frequency, and diabetic status. These parameters were explored using seven distinct machine learning approaches, namely Logistic Regression, Multilayer Perceptron, Sequential Minimal Optimization (SMO), Decision Stump, Hoeffding Tree, J48, and LMT. Our objective was to uncover hidden patterns and identify the most influential parameter for making future predictions. The results, along with accompanying numerical representations, are presented in Tables 1 to 8 and Figures 1 to 7.

Table 2 outlines the quality of the provided data by distinguishing correctly and incorrectly classified instances, while Table 3 expresses these distinctions as percentages. Most machine learning and deep learning approaches demonstrate superior accuracy in correctly classifying instances. Figure 1 and Figure 2 provide visual representations of these results.

We introduced Equation 1, which employs data from Table 4 and Figure 3 to calculate the Kappa statistics, a measure of inter-rater agreement or reliability, typically used to assess the consistency of ratings or classifications among multiple observers. Notably, the Decision Stump yields the lowest Kappa value, while the other machine learning approaches yield approximately 0.58, signifying better agreement. The deep learning approach for diabetic prediction returns a superior Kappa value of 0.89, as shown in Figure 3.

For error analysis, we employ the Mean Absolute Error (MAE) using Equation 2, and all seven machine learning algorithms demonstrate exceptional error performance, nearly approaching 0. Similarly, the Root Mean Square Error (RMSE), calculated with Equation 3, also returns excellent performance, with some machine learning approaches and deep learning models achieving a 0% error rate. Detailed numerical data can be found in Table 5 and Figure 4.

The Relative Absolute Error (RAE), defined by Equation 4, measures the accuracy of predictions in percentage terms. Among the seven ML classification algorithms examined, SMO exhibits the highest error rate, while the remaining six approaches perform with minimal error. Comparable results are found in the Relative Root Square Error (RRSE) analysis, extending to both machine learning and deep learning models. This information is presented in Table 6 and Figure 5.

The time taken for model development is a critical consideration, as discussed in Section 4.0. Table 7 and Figure 6 show that the Multilayer Perceptron and LMT require the most time for model creation, while Decision Stump, Hoeffding Tree, and J48 exhibit the shortest model-building durations. Logistic Regression and SMO also fall within the quicker model creation category. These observations are reflected in the associated visual representations.

In the context of Table 8 and Figure 7, it is evident that Decision Stump returns the lowest True Positive (TP) rate, while the other machine learning and deep learning approaches consistently achieve strong TP rates, nearly reaching 0. The False Positive (FP) rate, representing false alarms, is favorable for most approaches, except for LMT, and high precision is observed across all methods except for Decision Stump. Recall test statistics and other accuracy parameters also demonstrate robust performance for most machine learning and deep learning approaches, with the exception of Decision Stump.

5. Conclusion and Future Research

We have addressed the limitations of our model, which include considerations such as age, gender, family history of diabetes, and various lifestyle factors. Additionally, we have acknowledged potential computational constraints that may have influenced model development. In terms of future steps, we propose exploring additional data sources, investigating improved algorithms and hyperparameters, and fine-tuning the model to enhance its predictive performance. The ongoing advancements in diabetes research and treatment options are continuously emerging, with the ultimate goal of improving the well-being of individuals living with diabetes and reducing the overall societal impact of the disease.

References

[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/tigganeha4/diabetes-dataset-2019