Diabetes Diagnosis and Classification Using Feed Forward Neural Network Algorithm

Chintam Anusha
Computer Science and Engineering, Vignan's Institute of Information technology(A),
Visakhapatnam, Andhra Pradesh
India
anusha.rohini07@gmail.com

A. Sravani
Computer Science and Engineering, GITAM Institute of technology (Deemed to be University),
Visakhapatnam, Andhra Pradesh
India
sravani61sravz@gmail.com

M.A.Praveen
Computer Science and Engineering, Vignan's Institute of Information technology(A),
Visakhapatnam, Andhra Pradesh
India
anjaneyaprawin@gmail.com

Abstract

Diabetes mellitus (DM) is a persistent sickness that may cause numerous difficulties. Machine learning methods are used to analyze and classification of diabetes. The learning-based calculations play an important role in supporting dynamic in infection conclusion and expectation. In this work, conventional categorization algorithms and artificial neural networks are researched for the diabetes dataset. Likewise, different execution strategies with various angles are assessed for the Naive Bayes, K-nearest neighbour, decision trees, Extremely Randomized tree, radial basis function and multi-layer perceptron (MLP) algorithms. It upholds the patient's assessment that conceivably experiences the ill effects of diabetes later on. This paper gave that the feed-forward neural network algorithm multi-layer perceptron algorithm gives the most noteworthy expectation precision with the least Mean square error rate of 0.15. The multi-layer perceptron (MLP) gives the most reduced bogus negative rate and bogus positive rate with the most elevated region under the curve of 0.88.

Keywords

Machine learning algorithm, Diabetes prediction, MLP, classification of diabetes, neural network.

1. Introduction

DM - Diabetes Mellitus is one of the metabolic issues with improperly increased blood glucose levels. The carbons burned through will be transformed into a sort of sugar called glucose, and it will be delivered into the circulation system. Insulin, a chemical that helps move glucose from the blood to cells. With this constant condition, the pancreas will create next to zero insulin, and once in a while the delivered insulin won't be consumed by the cells; this is named insulin opposition (Lebovitz 1999). As of now, diabetes is viewed as one of the deadly infections across the world, and individuals are being influenced in a tremendous number. Almost 422 million individuals are diabetic patients, and about 1.6 million passing are credited to diabetes consistently. In the course of recent many years, the number of cases and the predominance of diabetes are consistently expendable (Marks et al. 2000). DM is named type 1, gestational diabetes and type 2. The condition where the pancreas will deliver practically zero insulin is type 1 diabetes. If the insulin isn't consumed by the cells or not created insufficient amount, it is alluded to as T2D (Rubaiat et al. 2018).

During pregnancy, a high glucose level, a high level of sugar would expand the danger of intricacies like hearing misfortune, vision misfortune, neuropathy, dementia, heart illnesses, sorrow, stroke, retinopathy, etc. Early recognition assumes a noticeable part in infection discovery. It is one of the essential reasons for cardiovascular sicknesses, and there is a huge need to help the clinical dynamic cycle. Numerous scientists in various clinical analyses have utilized different AI procedures (Quartuccio, et.al 2018).

The majority of the analysts regard clinical master frameworks, and there has been a lot of examination in this field. The clinical specialists and information examiners work together persistently to make this framework more exact and, subsequently, helpful, in actuality. Ongoing overviews by the WHO showed a gigantic expansion in the end ascribed to diabetes and in diabetic patients consistently. Along these lines, early determination of diabetes is a critical worry among specialists and clinical professionals (Karim, & Rahman, 2013). Multitudinous PC based location frameworks were planned and laid out for breaking down and expecting diabetes. The standard recognizing measure for diabetes sets aside time. By the by, with the ascent of AI, we can foster an answer for this serious issue (Saxena, 2021).

To precisely anticipate the issue, a decent method that can address the diabetes presence through input attributes is required. With a decent model and an exact recognition procedure, the finding can be made more productive. Given the forecast, clinical experts can imagine biomedical determination utilizing designing devices that can naturally adjust to any startling future conditions. A drawn-out expectation calculation can assume an imperative part in arranging and provisioning. Knowledge frameworks can learn or adjust and alter practical conditions in light of new encounters or changes in utilitarian connections (Kulkarni, 2017).

2. Related Work

Al-Zebari et al (2019).have analyzed the presentation of different machine learning methods for diabetes identification. MATLAB characterization student apparatus has been utilized in this paper including decision tree, Logistic regression-Nearest Neighbor, discriminant analysis, Support Vector Machine, and gathering students, their variations with 26 classifiers are thought of. The outcomes are assessed on a 10-crease cross-approval premise and normal arrangement exactness is considered for execution measures.

Pethunachiyar et.al. (2020) utilized SVM with dissimilar piece capacities for the grouping of diabetes. The reproduction model of the proposed framework incorporates 5 stages. In the wake of gathering the information, the choice cycle is done by redressing the blunders (irregularity in the information or missing qualities or wrong data). The information will be isolated into the testing dataset (30%) and preparing (70%) For effective expectation, the SVM strategy has been chosen, and a model has been assembled. Test information is applied to the model to make the expectation. The polynomial, direct and outspread bit based SVM has been carried out in this work. The disarray grid is utilized for ascertaining expectation exactness. To assess three portion capacities, the Receiver Operating Characteristic curve is utilized. Straight portion with SVM predicts all the more precisely contrasted with different bits

Kohli et al 2020. have applied different AI procedures on three distinctive illness datasets for sickness forecast. Highlight determination is completed by in reverse displaying utilizing the p-esteem test. The proposed model incorporates 4 stages: Initially, the dataset is investigated in a Python climate. During information munging, the missing qualities are supplanted with mean worth and mode an incentive for the constant variable and unmitigated variable, individually. The highlights are chosen carefully to work on the exhibition of the model. The ascribes are disposed of utilizing the regressive determination strategy (in light of p-esteem, it is killed). Subsequent to choosing the highlights, the model is refitted. Five calculations, including choice tree, strategic relapse, irregular backwoods, versatile boosting, and backing vector machine, were analyzed. The dataset has been partitioned into a test dataset (10%) and preparation set (90%) and In the future information munging, determination of highlights and model fitting advances can be computerized; pipeline structure for preprocessing information would further develop results.

Dey et al(2018) have fostered a web application utilizing Tensor flow for the fruitful expectation of diabetes. This proposed model requires patient information for effective finding, and the strategies like SVM, ANN (Artificial Neural Network), KNN and Naive Bayes are utilized to anticipate the illness. The dataset is isolated into two sections: preparing and testing the dataset. Preprocessing of information and information standardization would expand the precision of the model. Min Max Scaler standardization model is utilized to further develops precision.

Shanthi et al.(2019) proposed and fostered a model for diagnosing T2D utilizing the ELM (Extreme Learning Machine) strategy. The ELM numerical model has one secret layer feed-forward network, making arbitrary secret hubs. Boundaries are arbitrarily produced for the secret hubs at first. The following yield lattice is determined, and afterwards, the organization's ideal weight is given as the yield. From the qualities, input weight, and initiation works, the yield is gotten. The enactment capacities accessible are a three-sided premise, sine, hard-limit, and

sigmoid. This model helps clinical specialists in estimating T2D.

Maham Jahangir et al. (2017) presents a clever forecast structure that utilizes AutoMLP (programmed multi-facet perceptron) joined with an exception discovery technique. This strategy includes two phases: preprocessing of information with exception location following via preparing of AutoMLP. In the subsequent stage, it is utilized to group the information. Contrasted with different designs of the neural organization, AutoMLP gives higher precision. The ascribes like plasma glucose level, pulse, and the occasions pregnant are discovered to be more applicable.

Mohebbi et al. (2017) utilized CGM (Continuous Glucose Monitoring) signals for adherence identification in diabetic patients. A lot of signs were reenacted utilizing a T2D adjusted variant of the MVP (Medtronic Virtual Patient) model. Diverse grouping calculations were thought about by utilizing a complete matrix search. Strategic relapse, Multi-Layer Perceptron and Convolutional Neural Network (CNN) methods have been utilized in this work. CNN shows better execution in the arrangement.

3. Methodology

The Diabetes dataset from the Pima Indians Diabetes Database (Prasannavenkatesan et al. 2021) is taken for the prescient examination in this exploration work. The considered dataset was cleaned utilizing the information preprocessing and information cleaning procedures, then, at that point the came about dataset has been considered for a few tests over various characterization calculations. The Indians Diabetes Database contains the patient's subtleties with diabetes status that is Non-Diabetes and Diabetes. The essential patient's data is utilized to analyze and anticipate Diabetes Mellitus among the populace. The considered Diabetes Mellitus dataset contains 766 records. The dataset contains highlights of patients like 1) Plasma glucose fixation a 2 hours in an oral glucose resilience test, 2) Number of times pregnant, 3) Triceps skinfold thickness in mm, 4) Diastolic circulatory strain in mmHg, 5) Body mass record (weight in kg/(tallness in m)^2), 7) Diabetes family work, 8) Age in years, and 9) Class variable (0 or 1), 6) 2-Hour serum insulin in mug/ml, (Ruby et Al. 2020).

Table 1. Sample record of the cleaned dataset

#Pregnant	Glucose	BP	BMI	DPF	Age	Class
9	181	64	23.3	0.672	32	Diabetes
0	87	66	28.1	0.167	21	Non-Diabetes
3	139	40	43.1	2.288	33	Diabetes
4	120	74	25.6	0.201	30	Non-Diabetes
5	82	50	31	0.248	26	Diabetes
3	193	70	30.5	0.158	53	Diabetes
5	107	92	37.6	0.191	30	Non-Diabetes
11	171	74	38	0.537	34	Diabetes
12	1	80	27.1	1.441	57	Non-Diabetes
2	189	60	30.1	0.398	59	Diabetes

The information preprocessing and cleaning measure (information ascription mean method) eliminates the missing and anomaly information esteems from the dataset. The came about the dataset in the wake of prehandling is decreased to 722 records with three required pertinent highlights of patient subtleties. There are 722 patient subtleties in the dataset, out of which 463cases are in the class of 'Non-Diabetes,' and 248 cases are in the class of 'Diabetes' with 46records are missing required fundamental qualities. Six mathematical highlights from the dataset are taken as the info ascribes, and one element is considered as the yield property. The patient's data is introduced in Table.1.

This examination work investigation the forecast of diabetes of non-diabetes patients utilizing an alternate AI calculation. Distinctive grouping models are applied to the diabetes dataset, and its presentation as far as precision, blunder rates, and region bends are assessed. This work incorporates assessing Naive Bayes, KNN, Decision trees, Extra Trees, Radial basis function and MLP.

4. Results and Discussions:

This segment sums up the predicted results of the radial basis function, Naive Bayes, KNN, decision trees, extra trees and multi-layer perceptron algorithms. The k-fold cross-validation is one of the re-sampling systems used to approve the ML methods on the restricted data population. In this paper, the 'k' esteem is picked as so, it tends to be known as a 5-fold cross-validation resampling strategy. The 5-fold cross-validation technique plans to decrease the inclination (bias) of the forecast model. Normally, the presentation of the machine learning expectation algorithms is estimated by utilizing a few measurements dependent on the grouping calculation. In this work, the expected results are evaluated by using the metrics such as accuracy, root means square error (RMSE), mean square error (MSE), the recipient working trademark region under bend (ROC_AUC).

In this paper, the accuracy of the prediction (that is, regardless of whether the patient is diabetic or non-diabetic) of various machine algorithms is determined. Very order model has a different prediction accuracy based on its hyper parameters and a certain level of improvement over other forecast models. This work thinks about 80 % data set for preparing and 20 % of the information samples for testing classification algorithms. In this work, each model's accuracy is compared, and its expected results are summed up in table.2.

Table 2. Accuracy score of classifiers

S. No	Classifier	Accuracy	AU_ROC
1.	K-Neighbors Classifier (KNN)	71.7241	0.77
2.	Decision Tree (DT)	66.8965	0.64
3.	Naive Bayes	77.2413	0.85
4.	Extra Trees	72.4137	0.72
5.	Multi-Layer Perceptron (MLP)	80.6890	0.88
6.	Radial Basis Function	68.2758	0.70

The multi-facet perceptron calculation predicts the diabetes cases more precisely than other comparative. The multi-facet perceptron is the feed-forward counterfeit neural organization. The MLP measures the diabetes dataset utilizing non-direct enactment capacities with 3 layers of neurons: input, output layer and hidden. The dataset was tried with a few neuron esteems and arranged the given dataset into two types of classes as diabetes and non-diabetes patients with diminished blunders. It works by planning the given weighted contributions to the yield of every neuron among the test data and preparing information. For every information focuses dependent on the mistake esteems, the testing datasets are ordered. The end goal is that the MLP algorithm given a higher prediction rate than the remaining algorithms.

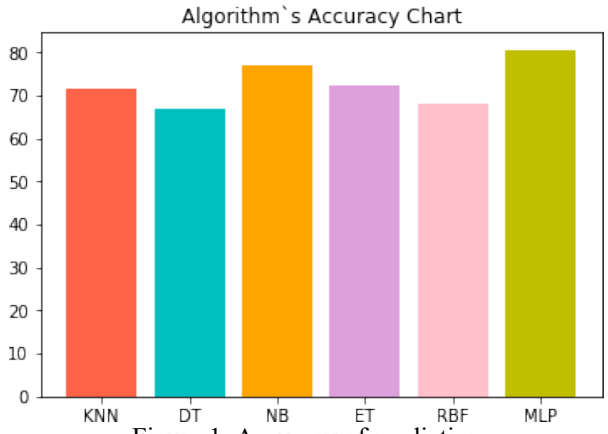

Figure 1. Accuracy of prediction

Figure 1 portrays the accuracy scores of various Classification Algorithms. From Figure.1, we can see that the multi-facet perceptron calculation has the most noteworthy exactness of 80.68. The multi-layer perceptron algorithm has 3.3 to 12.8 % of further developed exactness when contrasted with ET, KNN, NB, DT and RBF calculations. The multi-layer perceptron calculation works by characterizing the information point of the dataset of diabetes dependent on the comparability.

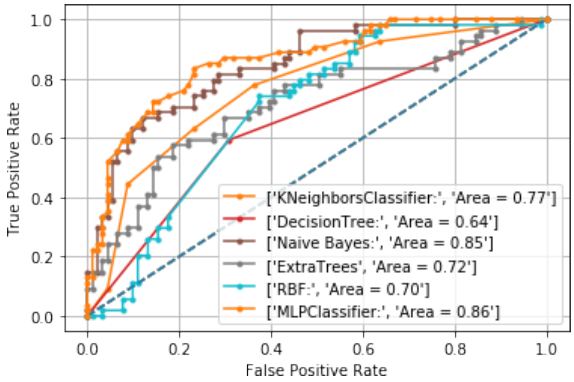

Figure 2. ROC_Accuracy Curve

Figure 2 given that portrayal between the true positive rate and false positive rate in the structure ROC region below the curve. The multi-facet perceptron calculation delivers the most elevated worth of 0.88 as contrasted and KNN, decision trees, Naive Bayes, extra trees, and radial basis function calculations.

Table 3 given the error measurements of the different ML methods. The bias measurements root mean square error and mean square error esteems for every calculation are assessed. The Naive Bayes, decision trees, radial bias function, KNN, extra trees, and multi-layer perceptron methods have the MSE mistake rate as 0.3310, 0.2827, 0.2758, 0.3172, 0.2275 and 0.1564, separate**ly.**

Figure 3. Given that the multi-layer perceptron calculation delivers the most reduced rate of error as 0.1564 on the forecast of precise diabetes cases than the other calculation. The multi-layer perceptron calculation arranges the testing dataset by yield planning between the existing and new test data. Therefore, it brings about lower bias rates. Additionally, the multi-layer perceptron RMSE error rate is likewise exceptionally low when contrasted with the mistake paces of DT (0.57), RBF (0.56) ET (0.52), NB (0.47) and KNN (0.53) calculations, as displayed in Table 3.

Table 3. Error metrics of classifiers

S. No	Classifier	MSE	RMSE
1.	K-Neighbors Classifier (KNN)	0.2827	0.5317
2.	Decision Tree (DT)	0.3310	0.5753
3.	Naive Bayes	0.2275	0.4770
4.	Extra Trees	0.2758	0.5252
5.	Radial Basis Function	0.3172	0.5632
6.	Multi-Layer Perceptron (MLP)	0.1564	0.4394

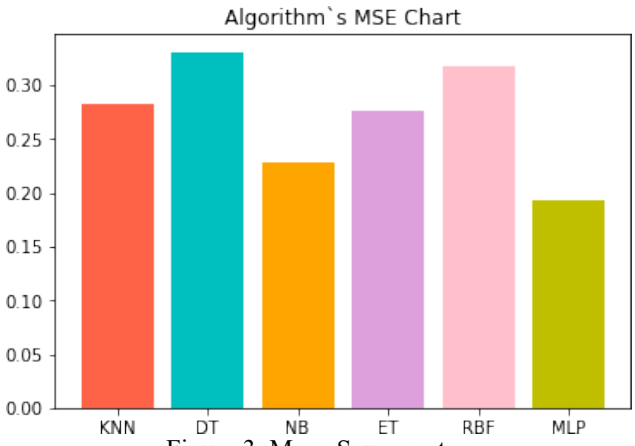

Figure 3. Mean Square rates

5. Conclusion

It is worth concentrating a lot of fundamentals to precisely foreseeing and diagnosing any infection by utilizing Machine learning. This work investigates distinctive ML Algorithms and the exhibitions on the dataset of diabetes. The consequences of ML calculations Naive Bayes, KNN, decision trees, extra trees and radial basis function are dissected in this study. All the above calculations have tried different things with the expectation accuracy, RMSE, MSE and ROC. The outcomes show that multi-layer perceptron has a superior exhibition with the diabetes dataset. Likewise, by contrasting and the consequences of other characterization calculations, we can see that the MLP has better ROC as 88% among Naïve Bayes, KNN, extra trees, radial basis function and decision trees. Subsequently, the outcomes propose that the MLP calculation can ready and characterize diabetic patients.

References

Lebovitz, H. E. (1999). Type 2 diabetes: an overview. *Clinical chemistry*, 45(8), 1339-1345.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Nelson, D. E., Engelgau, M. M., Vinicor, F., & Marks, J. S. (2000). *Diabetes trends in the US: 1990-1998*. Diabetes care, 23(9), 1278-1283.

Kalyani, R. R., Corriere, M. D., Donner, T. W., &Quartuccio, M. W. (2018). Diabetes Head to Toe: Everything You Need to Know about Diagnosis, Treatment, and Living with Diabetes. Johns Hopkins University Press.

Saxena, R. (2021). Role of K-nearest neighbour in detection of Diabetes Mellitus. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 373-376.

Karim, M., & Rahman, R. M. (2013). Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing.

Brownlee, J. (2020). How to Develop an Extra Trees Ensemble with Python. Machine Learning Mastery.

Kulkarni, M. (2017). Decision trees for classification: A machine learning algorithm. *The Xoriant*.

Orr, M. J. (1996). Introduction to radial basis function networks..

Plunkett, K., &Marchman, V. (2020). U-shaped learning and frequency effects in a multilayered perceptron: Implications for child language acquisition. *Connectionist psychology: A text with readings*, 487-526.

Adel Al-Zebari, AbdulkadirSengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection", 2019 *1st International Informatics and Software Engineering Conference* (UBMYK).

G. A .Pethunachiyar, "Classification Of Diabetes Patients Using Kernel Based Support Vector Machines", 2020 *International Conference on Computer Communication and Informatics (ICCCI -2020*), Jan. 22-24, 2020, Coimbatore, INDIA.

Pahulpreet Singh Kohli, Shriya Arora, "Application of Machine Learning in Disease Prediction", 2018 4th *International Conference on Computing Communication and Automation* (ICCCA).

Usha Ruby, A., Theerthagiri, P., Jeena Jacob, I., Vamsidhar, Y., "Binary cross entropy with deep learning technique for image classification"*International Journal of Advanced Trends in Computer Science and Engineering*, 2020, 9(4), pp. 5393-5397

S. Hari Krishnan, P. Vinupritha and D. Kathirvelu, "Non-Invasive Glucose Monitoring using Machine Learning", *International Conference on Communication and Signal Processing*, July 28 - 30, 2020, India.

M.Shanthi, RamalathaMarimuthu, S.N.Shivapriya, R.Navaneethakrishnan, "Diagnosis of Diabetes using an Extreme Learning Machine Algorithm based Model", 2019 *IEEE 10th International Conference on Awareness Science and Technology (iCAST).*

Sajratul Yakin Rubaiat, MdMonibor Rahman, Md.Kamrul Hasan, "Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection", 2018 *International Conference on Innovation in Engineering and Technology (ICIET).*

Prasannavenkatesan T, Jeena Jacob, I., Usha Ruby, A. and Yendapalli, V., 2021. Prediction of COVID-19 Possibilities using K-Nearest Neighbour Classification Algorithm*. Int J Cur Res Rev*| Vol, 13(06), p.156.

Ali Mohebbi, Tinna B. Aradottir, Alexander R. Johansen, "A Deep Learning Approach to Adherence Detection for Type 2 Diabetics", 2017 *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

Theerthagiri, P., "Forecasting hyponatremia in hospitalized patients using multilayer perceptron and multivariate linear regression techniques". *Concurrency and Computation: Practice and Experience*. Springer, 2021, e6248.

Maham Jahangir, Hammad Afzal, Mehreen Ahmed, KhawarKhurshid, Raheel Nawaz, "An Expert System for Diabetes Prediction using AutoTuned Multi-Layer Perceptron", *Intelligent Systems Conference* 20177-8 September 2017 | London, UK.

Sidong Wei, Xuejiao Zhao, Chunyan Miao, "A Comprehensive Exploration to the Machine Learning Techniques for Diabetes Identification", *2018 IEEE 4th World Forum on Internet of Things (WF-IoT).*

Prasannavenkatesan T, "Probable Forecasting of Epidemic COVID-19 in Using COCUDE Model", *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 26, e3, 2021, doi: 10.4108/eai.3-2-2021.168601.

Biographies

Ch. Anusha, working as an Assistant Professor in Vignan's Institute of Information Technology (A), Visakhapatnam, India. She has an experience of two years. She has published three paper in reputed international journals. Her research interests include computer networks, mobile ad-hoc networks, machine learning, Data mining.

A. Sravani, working as an Assistant Professor in Vignan's Institute of Information Technology (A), Visakhapatnam, India. She has an experience of seven years. Her research interests include Internet of Things, Data Mining and Image Processing. She published six papers in reputed journals out of that four are Scopus indexed. Also she authored a book on image processing.

M.A. Praveen, working as an Assistant Professor in Vignan's Institute of Information Technology (A), Visakhapatnam, India. He has an experience of one year. His research interests include computer networks, mobile ad-hoc networks, machine learning, Data mining.