International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 113-118, January-2024

https://doi.org/10.47001/IRJIET/2024.801015

Classification of Diabetes Using Machine Learning Technics

[1]Ihab Tareq Elias, [2]Muna M. Taher Jawhar

[1]Student, Department of Software, College of Computer Science and Mathematics, University of Mosul, Iraq
[2]Teacher, Department of Software, College of Computer Science and Mathematics, University of Mosul, Iraq

***Abstract* - Diabetes is one of the most widespread chronic diseases worldwide. More than 400 million people in the world suffer from it, and their number is increasing significantly every year. The continents of the world, especially Asia and Africa, are experiencing a high rate of infection due to a number of factors and environmental factors. Related to diet and healthy habits used in many developing countries. Diabetes is considered one of the most serious chronic diseases that affect the human body due to impaired secretion of the hormone insulin, because the pancreas is unable to secrete a sufficient amount of insulin, or when the cells of the human body do not accept insulin and don't use it. Effective, causing a sudden drop or rise in blood sugar levels. It has a huge impact on human health and type 2 diabetes is the most common form in humans as many factors contribute to its spread, such as genetic factors, poor diet and lack of regular exercise. In this paper we used the machine learning technical to classification of diabetes such as logistic regression, random forest, decision tree, vector classification, KNN and Naïve bayes and the accuracy was 77.66%, 76.33%, 98.66%, 75.33%, 80.66% and 80% The rest of the general standards used in classification are mentioned in the results of this paper.**

Keywords: diabetes, artificial intelligence, logistic regression, decision tree, vector classification, random forest.

I. INTRODUCTION

Public health is considered one of the important and necessary matters to protect society and protect it from diseases that are dangerous to health. Many countries spend a large amount of their gross domestic product for the well-being and happiness of their people, and initiatives and technologies such as vaccination have led to lengthening people's life expectancy [1]. However, the past several years have witnessed a significant emergence of chronic and hereditary diseases that affect public health. Diabetes is a common chronic disease that poses a major threat to human health. One of the characteristics of diabetes is that the level of glucose in the blood is higher than the normal level, occurring due to an imbalance in insulin secretion or a defect in its biological effects, or both [2].

Diabetes Mellitus - DM is a metabolic disorder that results in inappropriately high blood sugar levels. The carbohydrates consumed will be converted into a type of sugar called glucose, and insulin will be released into the bloodstream, which is the hormone that helps transport glucose from the blood into the cells. In this chronic condition, the pancreas produces little or no insulin, and sometimes the cells do not absorb the insulin produced, which is called insulin resistance [3]. Nowadays, diabetes is one of the deadly diseases all over the world, and people are afflicted by it in large numbers. About 422 million people suffer from diabetes, and about 1.6 million deaths are attributed to diabetes each year. Over the past few decades, the number of cases and prevalence of diabetes has increased steadily [4].

Accurate classification of diabetes is an essential step towards diabetes prevention and control in healthcare. However, early detection of diabetes is more beneficial in controlling diabetes. The process of identifying diabetes seems tedious at an early stage because the patient has to visit a doctor. Regularly, advances in machine learning methods have solved this critical and fundamental problem in healthcare by predicting disease. Many techniques have been proposed in the literature for diabetes prediction.

Machine learning is one of the branches of artificial intelligence. It focuses on creating systems that learn data and acquire knowledge to improve their performance automatically and without using programming directly. Machine learning relies on algorithms and models that allow systems to analyze data, gain experience, and make decisions, enabling them to adapt to tasks and improve their performance with time passing.

The rest of paper is structure as following: section 2 related works of classification of diabetes. Section 3, A brief description of machine learning classification techniques used in this paper. Section 4, the research methodology and discussion the result. Section 5 the conclusion.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 113-118, January-2024

https://doi.org/10.47001/IRJIET/2024.801015

II. RELATED WORKS

Many researchers have made contributions in the fields of diabetes prediction. Diabetes has a major economic impact on society.

In 2022, Cardozo (2022) addressed machine learning algorithms to help screen for diabetes through routine laboratory tests, using data from 62,496 patient laboratory tests. The following classifications were used: artificial neural networks, naive Bayes, K-nearest neighbor, random forest, models Regression and support vector machines in diabetes detection, artificial neural network model outperforms other models. Based on clinical data processing, computer processing has been used to identify diseases [5].

In 2018, Sisodia used deep learning methods on electrocardiogram (ECG) signals to detect diabetes. Specifically, convolutional neural network and long short-term memory were used by them and then the features were extracted by support vector machine. As a result, they found a very high accuracy of 95.7% [6].

In 2018, Wu et al. used three machine learning methods, i.e. decision tree (DT), naïve-based (NB) and support vector machine (SVM) on PIDD in order to predict diabetes. The Naïve Bayes classifier was found to be accurate by 76.30% [7].

In 2021, the author (Wu) and others sought to develop effective models for predicting early pregnancy diabetes mellitus (GDM). Seven variables and 73 covariate datasets were used to create models that predicted early GDM in different situations. In early pregnancy, ML models predicted GDM with high accuracy and were developed and tested in a Chinese population [8].

In 2019, Alehegn and others used random forests, KNN, NB, and J48 to develop diabetes analysis and prediction. The researchers used two datasets: PIDD (Pima Indian Diabetes Dataset) and the 130 American Hospital Diabetes Dataset. The developed system achieved 93.62 percent accuracy in the case of PIDD and 88.56 percent accuracy for a large set of data from 130 hospitals in the United States. For large dataset analysis, the NB and J48 prediction algorithms were found to be superior [9].

In 2019, Thomas et al conducted a study that implemented a decision tree algorithm to predict diabetes. Experiments were conducted on diabetes among the Pima Indians database, and the results achieved an accuracy of 87%. However, low sample sizes lead to poor precision. The systems developed can be used to predict or diagnose other patients in the same family [10].

In 2018, Zou et al compared four classification techniques, such as decision tree, ANN, logistic regression, and naive bay. More packing and boosting has been applied to everyone and the Random Forest has also been included. The maximum accuracy achieved by all was between 84% and 86% [11].

III. A BRIEF DESCRIPTION OF MACHINE LEARNING CLASSIFICATION TECHNIQUES

1) Support Vector Machine

It is a set of supervised and connected learning techniques used in classifications and regressions [12]. SVM can be described as a double-layer network where the given weights are nonlinear in principle and linear in the next layer. SVMs choose the constraints of the first layer as input vectors for training, as this leads to dimensionality reduction of Vapnik Chervonenkis (VC).

2) k-Nearest neighbor algorithm

K-Nearest Neighbor is the easiest machine learning algorithm based on supervised learning technique. The K-NN algorithm assumes the similarity between the new data case and the available cases and places the new case in the class that is most similar to the available classes.

The K-N-N algorithm saves all available data and classifies new data based on similarity. This means that when new data appears, it can be easily classified into a good group class using the K-NN algorithm. It can be used for regression as well as for classification but is mostly used for classification problems.

The working of K-NN can be explained on the basis of the following algorithm:[13].

Step 1: Select the K number of neighbors.
Step 2: Calculate the Euclidean distance for K neighbors according to Euclid's law shown in equation (1).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \qquad (1)$$

Step 3: Take the K nearest neighbors according to the calculated Euclidean distance.
Step 4: Among these neighbors, count the number of data points in each class.
Step 5: Assign new data points to that class whose number of neighbors is maximum.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 113-118, January-2024

https://doi.org/10.47001/IRJIET/2024.801015

3) Decision Tree

A decision tree algorithm, also known as a classification tree, is a type of supervising machine learning commonly used to address classification and regression problems. The classification tree, as the name suggests, is used to classify a data set into categories that belong to the target (response variable). If the target consists of categorical variables, then the type of decision tree is a classification tree. However, if the target consists of tactile or numerical variables, then the type of decision tree is a tree. Regression and solves numerical prediction problems [14].

A decision tree consists of a decision node, branches and leaves. The decision node represents the features of a data set that is divided into two or more homogeneous groups. The branches indicate the decision rules and the leaves indicate the outcome of the decision. The goal of using a decision tree is to create a training model that can be used to predict the class or value of a target variable by learning simple decision rules inferred from past data (training data). On the basis of comparison, we follow the branch corresponding to those values and move to the next node. Then the algorithm again compares the attributes with the child node until it reaches the leaves that indicate the decision result [15]. Decision tree algorithm [14].

The first step: Calculate entropy. Using Shannon entropy which in turn sorts the data set into homogeneous variables (by category) that have low information entropy and heterogeneous variables that have high information entropy.

Step Two: Calculate Average Information. Calculate the weight of the effect of each independent variable on the target variable using weighted entropy averages through equation.

Step Three: Calculate Information Gain. Calculate the information gain which is simply the difference between the entropy of the information found in Step 1 minus the average information calculated in Step 2.

Step Four: Create the Root Node. The independent variable with the highest information gain will become the root or first node on which the data set is divided.

Step Five: Create a Decision Tree. Repeat this process for each variable for which the Shannon entropy is nonzero. If the entropy of a variable is zero, then that variable becomes a "leaf" node.

4) Naïve Bayes

It is a supervised learning algorithm, based on Bayes' theory, used to solve classification problems and is mainly used in text that includes high-dimensional training data sets. Naïve Bayes Classifier is one of the simple and effective classification algorithms that helps in building fast learning models that can make fast predictions. It is a potential classifier, which means it is based on our potential object.

The Naïve Bayes algorithm is described as follows [16].

- Simple: Its sources depend on new principles, where there is a proportionality between all profit revenues independently of some people.
- Bayes: It is based on the principle of Bayes' theorem

His expectations are also given according to the rule of profit, which is likely to be imposed by long knowledge. It depends on the Canadian possibility.

5) Logistic Regression

Logistic regression is a supervised machine learning algorithm that is used for classification purposes. It is used when the data is in binary form, i.e. 0 and 1. Logistic regression predicts the output of the categorical dependent variable. Therefore, the result must be a categorical or discrete value. It can be either yes or no, 0 or 1, true or false, etc. But instead of giving the exact value as 0 and 1, it gives the probability values that lie between 0 and 1.

Logistic regression is very similar to linear regression except in how it is used. Linear regression is used to solve regression problems, while logistic regression is used to solve classification problems. Logistic regression uses the concept of predictive modeling as a regression; therefore, it is called logistic regression, but it is used to classify samples; Therefore, it falls within the classification algorithm. Logistic regression can be used to classify observations using different types of data and can easily identify the most effective variables used in the classification.

6) Random Forest

Breiman's robust learning algorithm (2001) proposed random forests, which is a well-known ensemble algorithm for developing prediction models that can be used to solve classification and regression problems. Random forests provide high classification performance and high generalization of data. The random forest classifier is a supervised machine learning method that creates a forest. First,it then combines the results of the trained decision trees using the bagging method (Polat, K., 2019).

The decision tree forms the basic classifier in the random forest, and randomization is done in two ways in creating random forests, one of which takes random samples to draw samples, and the other randomly selects attributes, or features; To create decision trees, decision trees are considered a good

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 1, pp 113-118, January-2024
https://doi.org/10.47001/IRJIET/2024.801015

candidate for classification, as they classify a large amount of data in terms of accuracy.

IV. RESEARCH METHODOLOGY

In this section we explain the analysis of the proposed system and how the designed system works and is a possible alternative to the current system. The data used in this paper was collected from the Kaggle dataset. The data has been subjected to different types of pre-processing, which will be discussed in subsequent sections to improve the system performance. The proposed model applies the classification model with the highest level of accuracy. These algorithms include decision tree, support vector machine classifiers and nearest neighbor. Structural diagram of the proposed model is illustrated in the following figure.

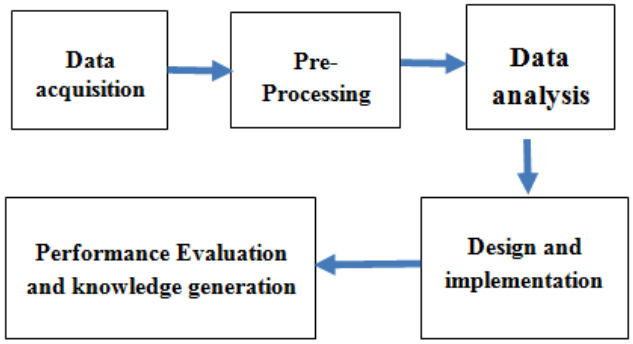

Figure 1: Structural diagram of the proposed model

1) Data acquisition

Initially, we downloaded the data from the Kaggle website. The goal of the data set is to make a diagnostic prediction of whether a patient suffers from diabetes or not based on some characteristics in the database (number of pregnancies, glucose test value, blood pressure, skin thickness, insulin, BMI, Age) are used to provide sufficient training data after pre-processing that requires removing certain entries.

2) Pre-processing stage

At this stage, we address discrepancies in the data to improve accuracy and results. This dataset contains missing values for some specific attributes such as glucose level and blood sugar because these attributes cannot have zero values.

The data value in different attributes contains some missing values. These missing values can lead to an inaccurate result, and may also reduce accuracy. Therefore, to deal with this missing value, the column method used to replace 0 with an appropriate calculation. To deal with these missing values programmatically, a package of Python was used to obtain an average function and process the existing values from 0 to the calculated ones. We also determine the priority of the attribute, so that the artificial neural network calculates Weight each neuron (feature) according to the selected priority. The priority feature is needed to have better accuracy in diabetes detection, which explains why priority-based diabetes detection is affected. Table 1 shows the attribute priority.

Table (1): Attribute priority

Attribute Name	Priority
Diastolic blood pressure (mm Hg)	1
Number of times pregnant	2
Age (years)	3
Triceps skin fold thickness (mm)	4
Diabetes pedigree function	5
Body mass index (weight in kg/(height in m)^2)	6
2-Hour serum insulin (mu U/ml)	7
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	8

The data set is then scaled to normalize all values. Correlation is the amount of context between properties. It is a real numerical value that indicates the degree of significance between 0 and 1. A negative value indicates an inverse relationship, while a direct relationship is indicated by a positive value. Figure 2 shows the correlation map of the proposed mode.

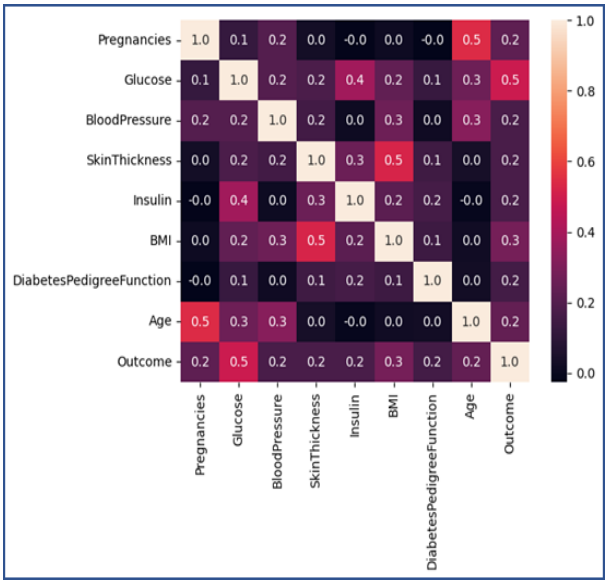

Figure 2: The correlation link map of the proposed mode

V. RESULT AND DISCUSSION

This stage is considered the last stage of building the model after all the stages have been clarified. To complete the process of building the system, we will present the results that we obtained, as the system was applied using the data set that

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 113-118, January-2024

https://doi.org/10.47001/IRJIET/2024.801015

was downloaded, in addition to a real database of people from whom information related to diabetes was taken, and then we carried out Testing the database. The results of the confusion matrix for all algorithms are shown in Table (2).

Table (2): Confusion matrix obtained by different methods

Logistic Regression			K-Nearest Neighbor			Decision Tree		
	0	1		0	1		0	1
0	168	52	0	160	35	0	179	0
1	15	65	1	23	82	1	4	117
SVC			**Random Forest**			**Naïve Bayes**		
	0	1		0	1		0	1
0	168	59	0	173	61	0	162	39
1	15	58	1	10	56	1	21	78

In the algorithms described above, the system was applied using individual methods separately to verify the results of the classification process. Table (3) shows the results of the initial implementation test for each method separately in combination with a training set size of 85% and a test set of 15% of the data set.

Table (3): Compare results of algorithms using individual methods

Method	Precision	F1	Recall	accuracy
Logistic Regression	55.5%	65.98%	81.25%	77.66%
Random Forest	47.86%	61.20%	84.84%	76.33%
Decision Tree	100%	98.31%	96.69%	98.66%
(SVM)	49.57%	61.05%	79.45%	75.33%
K-Nearest Neighbor	70.08%	73.87%	78.09%	80.66%
Naïve bayes	66.66%	72.22%	78.78%	80%

It turns out that the decision tree algorithm has the best results when using individual early diabetes classification algorithms. The difference between the algorithms is also explained in the diagram below.

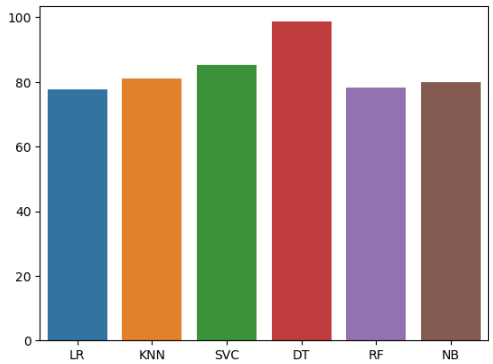

Figure (3): the difference between the accuracy of the algorithms

VI. CONCLUSION

One of the global health issues is to identify the risk of diabetes at its early phase. This study attempts to structure a framework which forecasts the risk pertaining to diabetes mellitus type 2. In this paper, six machine learning classification methods were implemented, and their results were compared with different statistical measures. Tests were performed on the dataset collected through online and offline questionnaires consisting 9 questions relevant to diabetes. Also, same algorithms were applied on PIMA database. The experimental result shows that the accuracy of decision six of our dataset is 98.66% which is the highest among the rest. decision tree is also giving highest accuracy for PIMA dataset.

Among three different machines learning algorithms applied, all the models produced good results for some parameter like precision, recall sensitivity etc. This study still holds a scope for further research and improvement including other machine learning algorithms to predict diabetes or any other disease.

REFERENCES

[1] Williams, R., Karuranga, S., Malanda, B., Saeedi, P., Basit, A., Besançon, S., Bommer, C., Esteghamati, A., Ogurtsova, K., & Zhang, P. (2020). Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Research and Clinical Practice, 162, 108072.

[2] Lonappan, A., Bindu, G., Thomas, V., Jacob, J., Rajasekaran, C., & Mathew, K. T. (2007). Diagnosis of diabetes mellitus using microwaves. Journal of Electromagnetic Waves and Applications, 21(10), 1393–1401.

[3] Lebovitz, H. E. (1999). Type 2 diabetes: an overview. Clinical Chemistry, 45(8), 1339–1345.

[4] 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.

[5] Cardozo, G., Pintarelli, G. B., Andreis, G. R., Lopes, A. C. W., & Marques, J. L. B. (2022). Use of Machine Learning and Routine Laboratory Tests for Diabetes Mellitus Screening. BioMed Research International, 2022.

[6] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585.

[7] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model

ISSN (online): 2581-3048
Volume 8, Issue 1, pp 113-118, January-2024
https://doi.org/10.47001/IRJIET/2024.801015

based on data mining. Informatics in Medicine Unlocked, 10, 100–107.

[8] Wu, Y.-T., Zhang, C.-J., Mol, B. W., Kawai, A., Li, C., Chen, L., Wang, Y., Sheng, J.-Z., Fan, J.-X., & Shi, Y. (2021). Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. The Journal of Clinical Endocrinology & Metabolism, 106(3), e1191–e1205.

[9] Alehegn, M., Joshi, R. R., & Mulay, P. (2019). Diabetes analysis and prediction using random forest,

[10] KNN, Naïve Bayes, and J48: An ensemble approach. International Journal of Scientific and Technology Research, 8(9), 1346–1354.

[11] Thomas, J., Joseph, A., Johnson, I., & Thomas, J. (2019). Machine learning approach for diabetes prediction. International Journal of Information, 8(2).

[12] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers in Genetics, 9(November),1–10.https://doi.org/10.3389/fgene.2018.00515

[13] Vapnik, V. (1998). Statistical learning theory. https://api.semanticscholar.org/CorpusID:61112307

[14] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.

[15] Kotu, V., & Deshpande, B. (2018). Data science: concepts and practice. Morgan Kaufmann.

[16] Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: an overview. Journal of Physics: Conference Series, 1142, 12012.

[17] Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. International Journal of Science and Research (IJSR), 5(1), 1842–1845.

[18] Polat, K. (2019). A hybrid approach to Parkinson disease classification using speech signal: the combination of smote and random forests. 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), 1–3.

AUTHOR'S BIOGRAPHY

Ihab Tareq Elias,
Student, Department of Software, College of Computer Science and Mathematics, University of Mosul, Iraq.

Citation of this Article:

Ihab Tareq Elias, Muna M. Taher Jawhar, "Classification of Diabetes Using Machine Learning Technics" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 1, pp 113-118, January 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.801015

****

