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

**ISSN (online): 2581-3048**

**Volume 8, Issue 1, pp 151-157, January-2024**

https://doi.org/10.47001/IRJIET/2024.801018

# Review 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 major diseases of the population across worlds. Diabetes is a chronic disease that occurs either when the pan- crease does not produce enough insulin or when the body cannot efficiently use the insulin it produces. Diabetes, one of the top 10 causes of death worldwide, is associated with the interaction between lifestyle, psychosocial, medical conditions, demographic, and genetic risk factors. Predicting type 2 diabetes is important for providing prognosis or diagnosis support to allied health professionals, and aiding in the development of an efficient and effective prevention plan. Several works proposed machine-learning algorithms to predict type 2 diabetes. However, each work uses different datasets and evaluation metrics for algorithms' evaluation, making it difficult to compare among them. 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. Machine learning algorithms have helped health professionals (including doctors) treat, analyze and diagnose medical problems, as well as detect disease patterns and other patient data. Machine learning can help people make an initial judgment about diabetes according to daily physical examination data, and can serve as a reference for doctors.**

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

## I. INTRODUCTION

Public health is considered one of the important and necessary concerns that protect society and protect it from health threatening diseases. Diabetes is a chronic disease that poses a serious threat to human health. Diabetes is currently one of the deadliest diseases worldwide, affecting a large number of people. Approximately 422 million people suffer from diabetes and approximately 1.6 million deaths are attributed to diabetes each year. In recent decades, the prevalence of diabetes has increased exponentially [1]. Diabetes Unit Sabah Al-Youm is the first step in preventing and detecting diabetes in healthcare settings. However, the most important issue in diabetes monitoring is early detection of diabetes. Renewing the method for determining diabetes is at some point laborious, so the patient has to consult a doctor.

In an effort to improve scientific knowledge to address this critical health problem through disease prediction, many diabetes prediction options have been developed in science.

Artificial intelligence is a field of computer science that deals with the simulation of intelligent behavior in computers. Artificial intelligence can be defined as the ability of a machine to simulate intelligent human behavior, as it enables a computer or computer-controlled robot to think intelligently and in the same way as humans People [2].

Overall, the field of artificial intelligence aims to advance the science and engineering of intelligence to create machines with human-like characteristics. This includes creating machines that can communicate, perceive, plan and represent knowledge, make decisions, move and manipulate objects, and learn from experience.

Machine Learning is an artificial intelligence specialist focused on developing systems that learn from data and gain specific insights to improve the data, and the application of programming directly benefits from this. Machine learning is based on algorithms and models that enable systems to analyze and collect a unique experience that can match their preferred algorithms and models. Depend Advances in machine learning techniques have led to applications that bring industrial benefits to humans, such as: E.g., robotic technology, search engines, information filtering, automation, medical diagnostics, autonomous vehicles, and robotic doctors [3].

Health's scientific research algorithms (including analytical) help analyze and diagnose medical problems, as well as identify the latest diseases and other patient data. learning can help people make an initial diabetes assessment based on daily data and can provide an accessible resource for clinicians [4].The essence of using a machine learning

 

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

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 151-157, January-2024

https://doi.org/10.47001/IRJIET/2024.801018

classifier and data mining is to extract knowledge from the information stored in the dataset and create a simple description of the model. A diabetes diagnostic tool should be developed that uses machine learning to predict diabetics and detect the disease early before it becomes pathological. Machine learning algorithms (MLA) identify patterns from statistical data sets and feed them to the system for digital processing. Great strides have been made in the field of using machine learning algorithms to solve many healthcare problems as technology advances. Some of them are used to diagnose and/or predict diabetes to make active and accurate decisions. Therefore, this study focuses on applying machine learning techniques to an online dataset to discover hidden patterns in medical diagnosis and prediction of diabetes using the collected data.

## II. RELATED WORK

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

In 2022, Cardozo 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 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 [6].

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 [7].

In 2021, authors Khaleel and Al-Bakry proposed a model that can predict whether a patient suffers from diabetes or not. This model relies on the prediction accuracy of powerful machine learning algorithms, which use certain metrics such as precision, recall, and F1 measure. The authors use the Pima Indian Diabetes Dataset (PIDD) to predict the onset of diabetes based on the diagnostic method. The results obtained using Logistic Regression (LR), Naïve Bayes (NB), and K-nearest Neighbor (KNN) algorithms were 94%, 79%, and 69%, respectively [8].

In 2017, Käräjämäki et al presented a unified framework for diabetes prediction based on machine learning. Six machine learning classifications for diabetes prediction and different evaluation criteria were used to investigate the performance of these classification techniques. The analysis results showed that Naïve Bayes achieved the highest performance than other classifiers, obtaining an F1 measure of 0.74 [9].

The authors Yuvaraj and Sripreethaa were used a new approach of machine learning algorithms applied in Hadoop-based clusters to predict diabetes. This approach is applied in the database of diabetes and gastrointestinal diseases of Pima Indians and the results obtained show that Machine learning algorithms produce the most accurate diabetes prediction [10].

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% [11].

In 2020, researchers Tripathi Kumar applied four machine learning algorithms, random forest, nearest neighbor, support vector machine, and linear discriminant analysis in predictive analysis of early-stage diabetes. High accuracy of up to 87.66% goes to the random forest classifier [12].

The table (1) illustrated previous studies in field of Diabetes classification using machine learning.

**Table 1: Result of Previous Studies in the Field of Diabetes classification using machine learning**

| No. | Researcher | Data | Algorithm | Metrics | Results | Notes |
|---|---|---|---|---|---|---|
| 1 | (Cardozo et al., 2022) | data from 62,496 patient laboratory tests | artificial neural networks, naive Bayes, K-nearest neighbor, random forest, models | computer processing | artificial neural network model outperforms other models | addressed machine learning algorithms to help screen for diabetes through routinem laboratory |

 

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 1, pp 151-157, January-2024
https://doi.org/10.47001/IRJIET/2024.801018

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | Regression and support vector machines | | | tests |
| 2 | (Alehegn et al., 2019) | used two datasets: PIDD (Pima Indian Diabetes Dataset) and the 130 American Hospital Diabetes Dataset | random forests, KNN, NB, and J48 | Accuracy | 93.62 in PIDD and 88.56 in 130 hospitals in the United States | Two databases were used |
| 3 | (Thomas et al., 2019) | Experiments were conducted on diabetes among the Pima Indians database | a decision tree algorithm | Accuracy | the results achieved an accuracy of 87% | low sample sizes lead to poor precision |
| 4 | (Khaleel & Al-Bakry, 2023) | The authors use the Pima Indian Diabetes Dataset (PIDD) | Logistic Regression (LR), Naïve Bayes (NB), and K-nearest Neighbor (KNN) | precision, recall, and F1 measure | Logistic Regression 94%, Naïve Bayes 79%, and K-nearest Neighbor 69%, | a model that can predict whether a patient suffers from diabetes or not |
| 5 | (Käräjämäki et al., 2017). | | Six machine learning classifications for diabetes prediction | F1 | F1 measure of 0.74 | used to investigate the performance of these classification techniques |
| 6 | (Yuvaraj & SriPreethaa, 2019) | This approach is applied in the database of diabetes and gastrointestinal diseases of Pima Indians | Machine learning algorithms | applied in Hadoop-based clusters to predict diabetes | the results obtained show that Machine learning algorithms produce the most accurate diabetes prediction | |
| 7 | % (Sisodia & Sisodia, 2018) | Sisodia used deep learning methods on electrocardiogram (ECG) signals to detect diabetes | neural network and long short-term memory were used by them and then the features were extracted by support vector machine | Accuracy | accuracy of 95.7% | used deep learning methods on electrocardiogram (ECG) signals to detect diabetes |

## III. MACHINE LEARNING ALGORITHM

As we all know, the risk of developing diabetes in the current era is increasing significantly, so it is desirable to know the different possible ways of classifying diabetes. In this research, it is proposed to use different machine learning algorithms to classify diabetes, and it is necessary to select the machine learning algorithm that will provide high accuracy in the type of data set. This comparative analysis will examine the efficiency of different machine learning algorithms and then determine which algorithm is best for which type of data, as we know that different machine learning algorithms classify early diabetes differently. As a result, determining the appropriate approach for a particular type of data set is crucial.

### 3.1 Support vector machine

It is a set of supervised and connected learning techniques used in classifications and regressions [13]. 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). The SVM network structure is defined in Figure (1).

 

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 1, pp 151-157, January-2024
https://doi.org/10.47001/IRJIET/2024.801018

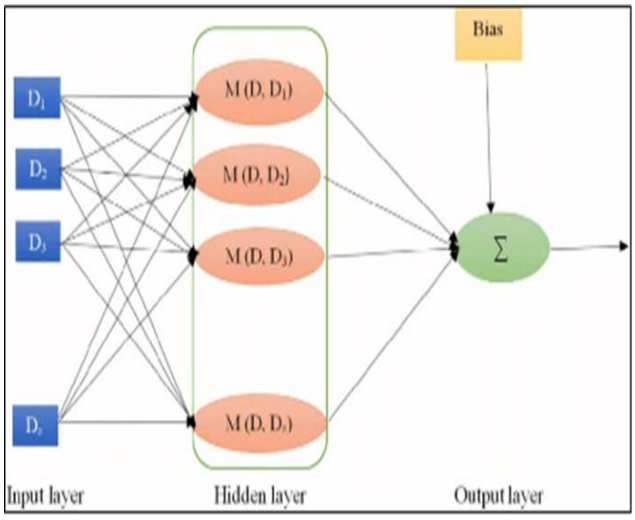


**Figure 1: SVM algorithm diagram**

The mechanism of the support vector machine algorithm is [14]:

Step 1: Find the boundaries of each category. Figure (1) SVM algorithm diagram
Step 2: Find the super level (a decision that separates two or more classes) where there are an infinite number of superlevels available.
Step 3: Once the limit and the hyperplane are defined, any new point can be classified by calculating on which side of the hyperplane it lies. This can be easily calculated by substituting the new test point x into the hyperplane equation. If the number +1 is calculated, it belongs to the class It is positive, and if it is calculated to -1, it belongs to the negative category.

### 3.2 K-Nearest Neighbor

The nearest neighbor algorithm (KNN) is one of the supervised learning algorithms that has been widely used in classification problems. The nearest neighbor algorithm is one of the most widely used classification methods due to its many interesting features, including its effectiveness and easy implementation [15].

The nearest neighbor algorithm is a basic and simple classification technique when there is little or no prior knowledge about the distribution of the data. In its work, the algorithm relies on measuring the Euclidean distance between each point and the point closest to it, and when the data is close to each other, the Euclidean distance is very small between each point and the point next to it, but as the data values become more distant and scattered, the distances between the points become large, and this is where the title of the algorithm came from, as the letter indicates K refers to the cases that will be classified based on the distances between them (i.e. between neighbors) [16]:

The working of K-NN can be explained on the basis of the following algorithm:

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)

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

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.

### 3.3 Naïve bayes

Naive Bayes classifier is one of the simple and effective classification algorithms that can be used to build fast machine learning models that can make predictions quickly. It is a probabilistic classifier, meaning it makes predictions based on the probability of an object appearing. The Naive Bayes algorithm (Naive Bayes) can be described as follows [17].

- Simple: Because it is based on the principle of independent possibilities, because it considers the relationship between all properties to be independent of each other.
- Bayes: Because it depends on the principle of Bayes' theorem Bayes Theorem is also known as Bayes Rule and is used to determine the likelihood of a hypothesis based on prior knowledge. It depends on the conditional probability.

Bayes' theorem is also known as Bayes' rule, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula of Bayes' theorem is given as follows:

$$P(\frac{A}{B}) = \frac{P(\frac{B}{A})P(A)}{P(B)} \quad (2)$$

P(A|B)) is the posterior probability: the probability of hypothesis A over the observed event B.
P(B|A)) is the probability of probability: the probability of evidence provided that the probability of the hypothesis is true
P(A) is the prior probability: the probability of the hypothesis before the evidence is observed.
P(B) is the prior probability of the forecast.

 

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

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 151-157, January-2024

https://doi.org/10.47001/IRJIET/2024.801018

Make a Naïve Bayes classifier:

The operation of the Naïve Bayes classifier can be understood based on the following steps:

Step 1: Calculate the prior probability for certain categories.
Step 2: Find the probability with each attribute for each category.
Step 3: Put this value into Bayes' Law and use the equation to find the posterior probability.
Step 4: Find out which category has the greatest probability, where the entry has the highest probability.

### 3.4 Logistic regression

Logistic regression is a supervised learning technique [18], which uses a logistic regression model to predict the probability of an event occurring by fitting data to a logistic curve. The independent variable in logistic regression can be quantitative or qualitative, and this is what distinguishes it from linear regression in which the independent variable is a continuous variable. Logistic regression is widely used in medical, social, agricultural and other fields. There are two types of logistic regression:

- Binary response logistic regression.
- Multivariable logistic regression.

Logistic regression is concerned with analyzing data with a binary response in which there is usually a dependent variable Preventively, in the case of success, the response variable takes the value (1) and in the case of failure, it takes the value (0) based on the Sigmoid function, which is also called the logistic function, as it gives a curve in the shape of the letter "S".[19-20]

### 3.5 Decision Tree

A decision tree algorithm, also called a classification tree, is a type of supervised machine learning often used to solve classification and regression problems. A decision tree is one of the most widely used data processing techniques for several reasons. The most important of these is that it is easy to prepare from the analyst's perspective and also easy to interpret for the user. Figure (2) shows a decision tree. As the name suggests, a classification tree is used to classify a data set into categories associated with the target (response variable). If the target is categorical variables, the decision tree type is a classification tree. However, if the target is tactile or numerical variables, the decision tree type is tree. Solving Regression Problems and Numerical Prediction [14]. A decision tree consists of a decision node, branches and leaves.

A decision node represents the characteristics of a data set divided into two or more homogeneous groups. The branches indicate the decision rules and the leaves indicate the result 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 derived from past data (training data). To predict a specific category of a data set, you start with the root node and the algorithm compares the root node with the actual features of the data set. Based on the comparison, we follow the branch corresponding to these values and move on to the next node. The algorithm then compares the attributes again with the child node until it reaches the leaves indicating the result of the decision [21]. The best feature in the decision tree is searched through certain sets of the following metrics to summarize each feature at each tree node [22].

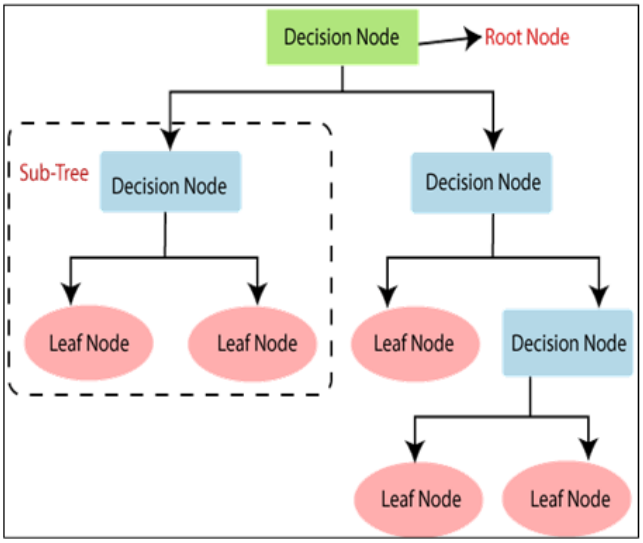


**Figure 2: Decision Tree algorithm diagram**

### 3.6 Random Forest

Robust Learning Algorithm Breimanin 2001 proposed Random forests, a well-known ensemble algorithm for developing predictive models that can be used to solve classification and regression problems. Random forests provide high classification performance and a high level of data generalization. Random Forest Classifier is a supervised machine learning method that creates a forest. First, it combines the results of decision trees constructed using the bagging method [23]. The decision tree is the basic classifier in random forests. Randomization occurs in two ways when creating random forests: one randomly selects samples to select samples, and the other randomly selects attributes or features. When it comes to creating decision trees, decision trees are considered good candidates for classification because they classify a large amount of data in terms of accuracy. Figure (3) shows the random forest algorithm.

 

ISSN (online): 2581-3048
Volume 8, Issue 1, pp 151-157, January-2024
https://doi.org/10.47001/IRJIET/2024.801018

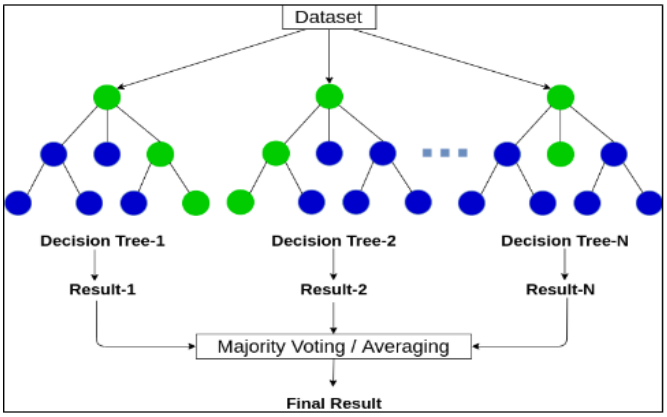


**Figure 3: Structure of the random forest algorithm**

## IV. CONCLUSION

Diabetes is a disease that can cause many complications. Many people in the world suffer from this disease, which often leads to death. The causes that lead to diabetes vary, including genetic and pathological, but early diagnosis of the disease leads to a very high probability of treatment and avoiding complications.

There are many methods used to diagnose and predict diseases, including artificial intelligence. Many researchers have used machine learning and deep learning techniques to classify and detect various diseases. In this study, we focused on detecting and classifying diabetes, the techniques used, statistics resulting from previous research, and the techniques used to detect diabetes.

How to accurately predict and diagnose this disease using machine learning is worth studying. In addition, by comparing the results of six classifications, we can find that there is no significant difference between random forest, decision tree and logistic regression, but decision tree is obviously better than other classifiers in some methods. This could indicate that machine learning can be used to predict diabetes.

## REFERENCES

[1] 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.

[2] Sindu, M. (2018). ARTIFICIAL INTELLIGENCE VS HUMAN RESOURCE PRACTICES IN BANKING SECTOR. International Journal of Social Sciences and Economic Research, 3(9), 5152–5258.

[3] Zhang, B., Lu, L., & Hou, J. (2019). A comparison of logistic regression, random forest models in predicting the risk of diabetes. Proceedings of the Third International Symposium on Image Computing and Digital Medicine, 231–234.

[4] Lee, B. J., & Kim, J. Y. (2015). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE Journal of Biomedical and Health Informatics, 20(1), 39–46.

[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] Alehegn, M., Joshi, R. R., & Mulay, P. (2019). Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An ensemble approach. International Journal of Scientific and Technology Research, 8(9), 1346–1354.

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

[8] Khaleel, F. A., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings, 80, 3200–3203.

[9] Käräjämäki, A. J., Bloigu, R., Kauma, H., Kesäniemi, Y. A., Koivurova, O.-P., Perkiömäki, J., Huikuri, H., &Ukkola, O. (2017). Non-alcoholic fatty liver disease with and without metabolic syndrome: different long-term outcomes. Metabolism, 66, 55–63.

[10] Yuvaraj, N., &SriPreethaa, K. R. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Computing, 22(Suppl 1), 1–9.

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

[12] Tripathi, G., & Kumar, R. (2020). Early prediction of diabetes mellitus using machine learning. 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 1009–1014.

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

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

[15] Jiang, L., Cai, Z., Wang, D., & Jiang, S. (2007). Survey of improving k-nearest-neighbor for classification. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), 1, 679–683.

[16] Imandoust, S. B., &Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications, 3(5), 605–610.

 

ISSN (online): 2581-3048
Volume 8, Issue 1, pp 151-157, January-2024
https://doi.org/10.47001/IRJIET/2024.801018

[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] Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. B, 4, 51–62.
[19] Bisong, E. (2019). Building machine learning and deep learning models on Google cloud platform. Springer.
[20] Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019). Logistic regression model optimization and case analysis. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 135–139.
[21] Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: an overview. Journal of Physics: Conference Series, 1142, 12012.
[22] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20–28.
[23] 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.
Email:
ihab.22csp9@student.uomosul.edu.iq

**Citation of this Article:**

Ihab Tareq Elias, Muna M. Taher Jawhar, "Review Classification of Diabetes Using Machine Learning Technics" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 1, pp 151-157, January 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.801018

*******

 