Review on Prediction of Diabetes Using Machine Learning Classification Algorithms

Priyanka Sopan Mogal[1], Prof. K D kharat[2]

[1]CSMSS, Chatrpati Shahu College of Engineering, Kanchanwadi Chh. Sambhajinagar
[2]CSMSS, Chatrpati Shahu College of Engineering, Kanchanwadi Chh. Sambhajinagar

---***---

Abstract - Diabetes is a global health situation that can impact a patient's whole life. Diabetes affects people of every age group. Developing technology offers a novel technique to predict diabetes and improve efficiency. Researchers mostly take the Pima Indian dataset to predict diabetes. In this study, the authors construct a structure, which can accurately evaluate the chance of diabetes in patients. This research presents a framework for accurately estimating the chance of diabetes in patients. The authors propose using Machine Learning techniques like Decision Trees, SVM, and Naive Bayes to diagnose diabetes in its early stages. The Pima Indian Diabetes test from the UCI library is utilized to predict diabetes, saving time and ensuring accuracy.

***Key Words*:** Machine Learning, Decision Tree, Naïve Bayes, Diabetes, Precision, Classification, SVM.

1. INTRODUCTION

Diabetes is a well-known condition characterized by impaired insulin production and response, leading to elevated blood glucose levels.

Diabetes may be classified into two types: type 1 (T1D) and type 2 (T2D). Type 1 diabetes often affects individuals under 30 years old. Clinical signs include increased thirst and frequent urination, as well as elevated blood glucose. Patients with this kind of diabetes require insulin therapy as oral medications solely are ineffective. Type 2 diabetes is more common in middle-aged and elderly people, and is often associated with obesity, hypertension, a high level of arteriosclerosis, and other disorders.

According to the World Health Organization (WHO), diabetes affects approximately 422 million people, with a higher prevalence in low-income countries. Additionally, this figure has the potential to increase to $490 billion by 2030. Diabetes affects several countries, including China, Canada, and India. Diabetes is becoming one of the leading causes of mortality globally. Early monitoring of diseases like diabetes can save lives.

Machine learning is used to forecast or diagnose life-threatening illnesses, including cancer, diabetes, heart disease, and thyroid. This study aims to predict diabetes by analyzing many factors associated with the disease.

We use the Pima Indian Diabetes dataset and various Machine Learning (ML) techniques to predict diabetes. Machine learning (ML) is a method for expressing instructions to computers or machines. ML algorithms may effectively acquire knowledge by creating grouping and classification models from a dataset. Choosing a suitable machine learning algorithm for predicting might be challenging. This study examines the efficiency of several algorithms for prediction diabetes, including K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT).

2. Literature Review

Deepti Sisodiaa and Dlip Singh Sisodiab[1] present their work on diabetes prediction at an early stage. Prediction algorithms include decision trees, SVMs, and Naive Bayes. The accuracy is assessed using the WEKA tool. Naïve Bayes demonstrated the highest accuracy.

Xue-hui Menget et al. [2] conducted a comparison of three algorithms for predicting diabetes or prediabetes based on common risk variables. The logistic, ANN, and decision tree methods are compared. The 735 patients that were tested were from two areas in Guangzhou, China. The maximum degree of accuracy (77.87%) is reached by the classification algorithm.

V. Anuja Kumari et al. in [3].The classifier uses SVM and the Pima Indian Diabetes dataset to predict diabetes illness at the lowest possible cost and with superior performance. Matlab R2010a was used for the implementation. The problem's accuracy results in 78.5%.

Monisha.A et al. [4] employed several classifiers in machine learning to predict and diagnose diabetes. Examples include Naive Bayes statistical modeling, logistic regression, and Extreme Gradient Boosting. Diabetes datasets for Pima Indians are being explored. Extreme Gradient Boosting method offers an accuracy rating of 81%, which is higher than the other two algorithms.

S. Selvakumar et al. [5] discussed diabetic problems. Data mining technologies are utilized to predict if a person is diabetic or not. Binary Logistic Regression, Multilayer Perception, and K-Nearest Neighbour methods are categorized. The accuracy level for Binary Logistic Regression is 0.69, Multilayer Perception is 0.71, and K-Nearest Neighbour is 0.80. K-Nearest Neighbour has higher accuracy than Binary Logistic Regression and Multilayer Perception.

According to Aiswarya Iyar et al. [6], diabetes affects 246 million individuals globally. According to WHO, these numbers will climb by more than 380 million by 2025.The purpose of this study is to discover a way to diagnose the condition. Using the decision tree and Naive Bayes algorithms. The Weka tool is utilized for implementation. The naive Bayes method achieved 79.5652% accuracy.

B.Tamilvanan et al. in [7] this paper's purpose is to predict diabetes with more accuracy. The accuracy rates of three categorization methods are compared: Naive Bayes, Random Forest, and NB-Tree. Implementation using the Weka tool. Naive Bayes has the highest accuracy rate (76.3%) and the lowest error rate (23.7%), making it the most effective prediction model.

Rahul Joshi et al. [8] used machine learning approaches to forecast medical datasets at an early stage, which is safe for human life. To evaluate the Pima Indians' diabetes dataset. The algorithms used include KNN, Naive Bayes, Random Forest, and J48. The ensemble approach yields the greatest results when individual approaches and procedures are combined. It is also termed a hybrid model. This delivers superior performance and accuracy over the single one. Weka and Java technologies are used to forecast diabetes.

Amina Azar et al. [9] Diabetes affects both young and old peoples. These are becoming more common by the day, and there is no cure. Data mining is used to make early-stage predictions. This paper's major goal is to differentiate and recommend the best algorithm. The PID datasets are utilized. The Decision Tree, Naive Bayes, and K-Nearest Neighbour algorithms are examined and utilized to predict diabetes diagnoses at an early stage with the greatest accuracy and efficiency. WEKA is used for testing and validation of fast miner. As a consequence, the decision tree is the optimal prediction method. It offers an accuracy level of 75.65%.

Veena Vijayan.V et al. in [10]. Choosing proper classification algorithms significantly improves the system's accuracy and efficiency. The major goal of this work is to analyze the benefits of several pre-processing strategies for decision support systems for predicting diabetes that are based on Support Vector Machine (SVM) and Naive Bayes classifier. The pre-processing methods employed in this study were Principal Component Analysis and Discretisation. The accuracy variation was assessed with and without pre-processing procedures. The Weka tool was utilized in this investigation. The dataset was obtained from the University of California, Irvine's (UCI) machine learning library.

DeepikaVermaet al. in their paper [11] uses two disease datasets. That is a breast cancer and diabetes dataset from the UCI machine learning repository. This work uses the WEKA tool, which is a good categorization tool. To classify a dataset of breast cancer and diabetes using Naive Bayes, SMO, REP tree, J48, and MLP algorithms using the WEKA interface. After examining the performance of all algorithms, J48 achieves 74.28% accuracy level than the other algorithms on the breast cancer dataset, whereas SMO obtains 76.80% accuracy level on diabetes.

3. Methodology

This effort aims to predict diabetes using several machine learning approaches and compare the results to choose the most accurate classifier. In the accompanying, we briefly discuss the steps. Figure 1 shows the flowchart of the suggested model for diabetes prediction.

3.1. Dataset Description

This article uses the Pima Indian Diabetes Dataset. The University of California, Irvine AI respiratory dataset (P.I.D.) is open and accessible. This dataset has 768 records and nine characteristics, including the result attribute. Out of 768 reports, 268 are "tested positive," indicating diabetes, while 500 are "tested negative," indicating the patient does not have it.

1 Polyuria: the term polyuria refers to the excessive production or passage of urine.
Diabetes is caused by excessive blood glucose levels. When blood glucose levels reach the renal threshold, the kidneys are unable to reabsorb all of the glucose, which spills into the urine.

2 Polydipsia: is the term used to describe excessive thirst aura increase fluid intake and it is a classic symptom of diabetes mellitus. In this glucose spills to the urine. Body loses large amount of water leading to dehydration

3 Sudden weight loss: is early common symptom of uncontrolled diabetes especially in type 1 diabetes when blood glucose level are high but insulin is either absent or not working properly the body cannot use glucose for energy. Instead it starts to break down fat and muscle to meet its energy need

4 Weakness: it is a diabetes result from poor glucose utilization, dehydration, and muscle loss aura complication like nerve & kidney damage. It is common but can usually be manage with good diabetes control.

5 Polyphagia: is excessive hunger caused by the body inability to use glucose for energy leading to cellular starvation despite high blood sugar. In this glucose build up in the blood because it cannot enter cells. Despite the fact that the blood has lots of glucose, cells are starved for energy.

The brain interprets these as hunger and saying signal to eat more.

6 Genital thrush: is a fungal infection usually caused by candida albicans an over growth of yeast in the genital area
In women: it affects the vagina and vulva
In men: it affects the head of the penis especially in uncircumcised men

7 Visual blurring: blurry vision is a common symptom of diabetes and it can occur suddenly or gradually.
It might be transient or the result of a significant diabetic eye problem.
It may cause by high blood sugar diabetic retinopathy muscle edema cataract glaucoma some causes are reversible but long term damage can lead to permanent vision loss of untreated.

8 Itching: it common but often overlook symptom especially when blood sugar level are poorly control. It can be generalized all over the body or localised commonly in the leg arm feet and genital area. Common causes are dry skin yeast infection itching in intense poor blood circulation nerve damage or allergic reaction.

9 Irritability: is a common emotional symptom and it often linked to blood sugar fluctuation as a result of blood sugar swings both high and low emotional stress or burnout from managing a chronic condition.

10 Delayed healing: is due to high blood sugar impairing circulation immune function and tissue repair even small cuts blisters or sores take much longer to heel increasing the risk of infection or ulcer particularly on the feet and legs.

11 Partial paresis: Partial paresis is muscular weakness or a lack of voluntary movement that can occur as a result of diabetes.
Is a form of nerve related muscle weakness caused by diabetes neuropathy. It can affect limbs facial muscles or internal organ like stomach.

12 Muscle stiffness: it may affect the hand shoulder legs or joint and is often related to complications of high blood sugar affecting nerve joint and connective tissues. Regular exercise stretching and good blood sugar control can help manage or prevent stiffness.

13 Alopecia: it refer to hair loss it is known as complication in people with poorly control long standing diabetes.
It may present as diffuse hair thinning patchy hair loss or slower hair regrowth.

14 Obesity: is a major risk factor and is closely link to insulin resistance.
Obesity might make blood sugar management more difficult and raise the likelihood of complications.

3.2. Data Preprocessing

The most important procedure is data preparation. This procedure is crucial for accurate and efficient predictions when using ML algorithms on a dataset.

The Indian Diabetes dataset has no missing values (NAN), however several zero-valued attributes are useless.

We calculate the mean and median of zero-valued columns for diabetes and non-diabetic patients. We substitute zero values for diabetic and non-diabetic patients.

The Pima Indians Diabetes dataset was normalized using the suggested methods, with 70% utilized for validation and training and 30% for testing.

3.3. Algorithms used for Classification

Once our dataset is prepared, we apply Machine Learning to classify it. This paper uses KNN, RF, SVM, ANN, and DT classification algorithms with dataset features such as pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, pedigree, and age. We also extract two additional features from the dataset using Exploratory Data Analysis. A diabetic is defined as having a blood pressure of more than 80 and a glucose level of more than 105. Diabetics have a blood pressure of 80 or above. The RF classification method achieves the greatest accuracy of 88.31% while considering the stated factors.

4. System Design

Classification involves categorizing data into specific classes and assigning labels to each one. Prediction models forecast continuous-valued functions. Data mining and machine learning approaches are utilized to effectively diagnose these issues.

Classification is an important approach for illness prediction. Classification is one of the most popular data mining jobs. Classification is common in large corporate and medical datasets. Classification is a data mining function that distributes objects in a collection to certain categories. Classifying diabetic patient datasets achieves the expected accuracy. Examples include J48, SVM, Naive Bayes, Decision Tree, logistic regression, and artificial neural networks (ANN). It is preferable to diagnose various disorders.

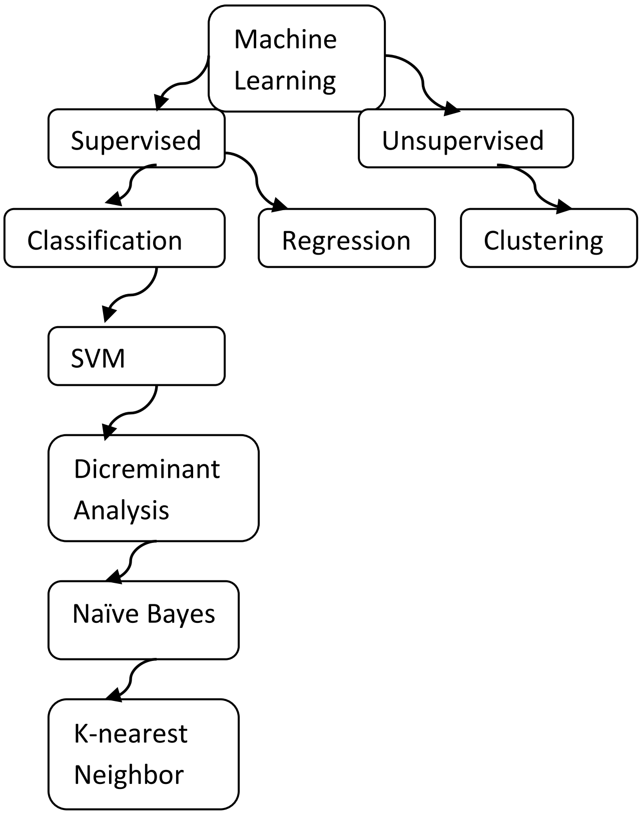

Fig -1: Flow Of Machine Learning

Classification is a key decision-making approach for real-world problems. The suggested method involves machine-learning techniques. The goal of this effort is to increase the accuracy of data categorization into diabetes or non-diabetic categories. Choosing more samples for classification problems does not necessarily result in improved accuracy. Although the algorithm performs well in terms of speed, it has low accuracy in data categorization.

The primary goal of our model is to obtain great accuracy. Using a large training dataset and a small testing dataset can improve classification accuracy. This study investigated several categorization approaches for identifying diabetic and non-diabetic individuals.

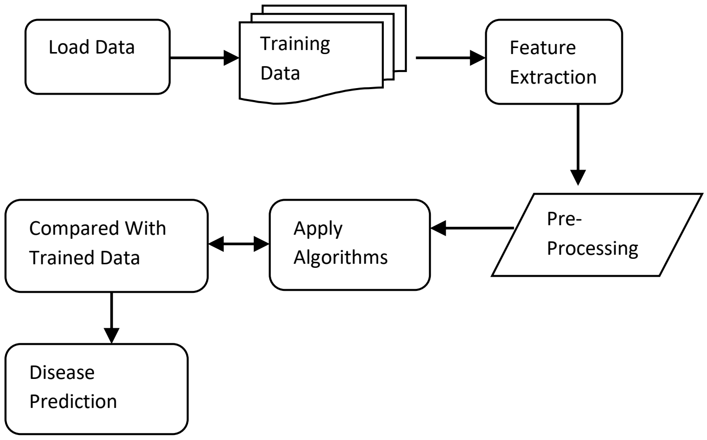

Fig -2: Fundamental Architecture for Prediction of disease

5. CONCLUSIONS

Machine learning, a type of artificial intelligence, has the potential to significantly improve diabetes risk prediction and early detection. Early diagnosis is crucial for successful diabetes management.

This paper aimed to predict diabetes using several machine learning approaches and analyze the results to identify the most accurate classifier, which we successfully did. To obtain high accuracy, we extracted two new features from the data set and explored different machine learning classification approaches. RF and ANN algorithms outperform other ML classification approaches in terms of efficiency and accuracy.

REFERENCES

1. DeeptiSisodia, Dilip Singh Sisodia,"Prediction of Diabetes Using Classification Algorithm", www.elsevier.com/locate/procedia, Procedia computer science 132(2018) 1578-1585.
2. Xue-Hui Meng,Yi-Xiang Huang,Dong-PingRao,Qiug Liu,2013,"Comparison of Three Data Mining Models For Predicting Diabetes of Prediabetes By Rick Factos",Kaohsiung journal of medical science(2013) 29,93-99.
3. V.AnujaKumari, R.Chithra."Classification of Diabetes Disease Using Support Vector Machine".vol 3.,Issue 2,March-April 2013,pp.1797-1801.www.ijera.com.
4. Monisha.A, S.ShalinChistina, Nirmala Santiago, "Decision support system for a chronic disease-Diabetes". International Journal of Computer &Mathematical Science (IJCMS), ISSN 2347-8527 Volume 7, Issue 3, March 2018.
5.S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar, "Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques", International Journal of statistics and Systems,ISSN 0973- 2675 Volume 12,Number 2(2017),PP.183-188.http://www.ripublication.com.
6. Aiswarya Iyar, S. Jeyalatha and RonakSumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
7. B.Tamilvanan, Dr.V.MuraliBhaskaran, "An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 39-44, www.iosrjournals.org.
8. Rahul Joshi, MinyechilAlehegn,"Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach",International Research Journal of Engineering and Technology (IRJET),Volume: 04 Issue: 10 | Oct -2017,e-ISSN: 2395-0056,p-ISSN: 2395-0072. www.irjet.net.
9. Amina Azar,Yasir Ali, Muhammad Awais, KhurramZaheer,"Data Mining Models Comparison for Diabetes Prediction", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 8, 2018.
10. VeenaVijayan.V, Anjali.C,"Decision Support Systems for Predicting Diabetes Mellitus –A Review", Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015), 978-1-4799-8553-1/15/$31.00 © s2015 IEEE.
11. DeepikaVerma , Dr.Nidhi Mishra, "Analysis and prediction of breast cancer and diabetes disease dataset using data mining classification techniques",Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017),IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9.