e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

FINDING THE HIGHEST ACCURACY USING MACHINE LEARNING ALGORITHMS IN DETECTION OF DIABETES

Amisha Mehta[*1]

[*1]Department Of Information Technology, B.K Birla College Of Arts, Science And Commerce, Mumbai, Maharashtra, India.

ABSTRACT

Diabetes mellitus is among the most dangerous disorders that a large number of individuals suffer from. Numerous causes, such as advanced age, obesity, poor diet, high blood pressure, genetic diabetes, and inactivity, can lead to diabetes mellitus. Diabetes raises a person's risk of heart disease, kidney disease, stroke, nerve damage, vision issues, and other conditions.

Hospitals use various tests to diagnose diabetes and treat patients accordingly. Big data analytics are essential in healthcare due to numerous databases. It helps analyze large datasets, uncover hidden patterns, and make accurate predictions. However, current diabetes classification and prediction accuracy are subpar. This study created a better diabetes prediction model by including additional traits beyond standard parameters like age, glucose, BMI, and insulin.

Keywords: Diabetes Mellitus, Big Data, Analysis, Machine Learning, Diabetes Detection.

I. INTRODUCTION

Diabetes is classified into three major types: Type 1 diabetes, also known as insulin-dependent diabetes mellitus (IDDM), is caused by the body's inability to produce enough insulin, requiring patients to take insulin injections. Type 2 diabetes, also known as non-insulin-dependent diabetic mellitus (NIDDM), is caused by the body's cells misusing insulin. Gestational diabetes (Type 3) is characterized by high blood sugar levels during pregnancy without a prior diabetes diagnosis. These diseases are linked to a variety of long-term health issues, emphasizing the importance of effective management.

Furthermore, people with diabetes are at a higher risk of developing numerous health issues, demanding close monitoring and research. Predictive analysis emerges as a critical tool for addressing these difficulties. It combines machine learning algorithms, data mining approaches, and other technologies.

Using statistical tools to glean useful insights and forecast future healthcare events, this analytical technique enables healthcare practitioners and researchers to make educated decisions and exact forecasts, allowing for improved patient care, more effective management, and the progress of diabetes research.

Machine learning and regression techniques can be used to perform predictive analytics. Predictive analytics strives to improve clinical outcomes by accurately diagnosing diseases, improving patient care, optimizing resources, and optimizing resources.[1]

II. METHODOLOGY

A diabetes dataset is taken and analysis is done on the basis of that dataset. Feature matching of the data to be learned using existing methods is part of the challenge of selecting a machine learning algorithm.

Architecture

Figure 1. is the architecture diagram for the diabetes prediction model. This model has five different modules. These modules include-

1. Dataset Collection
2. Data Pre-processing
3. Clustering
4. Build Model
5. Evaluation

This is the methodology that is going to be followed.

III. MODELING AND ANALYSIS

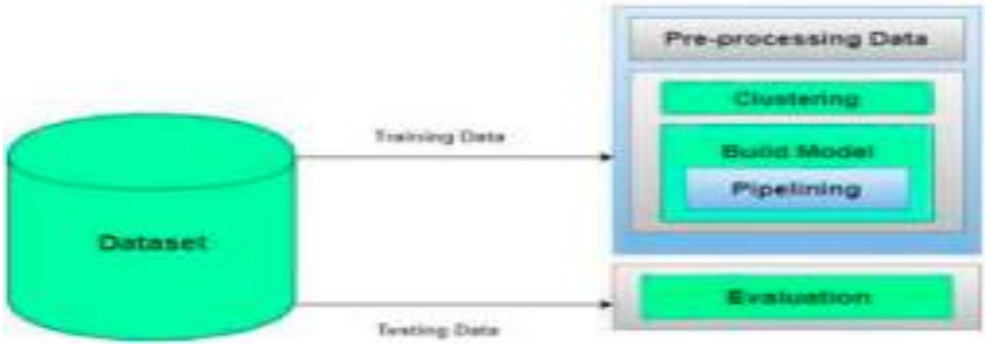


Figure 1: Diabetes Prediction Model

According to Figure 1. and the modules that were described in the previous section, we will study each module in detail.

1. Dataset Collection

This module includes data collection to study data patterns and trends which are helpful in prediction and evaluation of the results. Dataset description is given below:

Table 1:

Attributes	Type
Number of Pregnancies	N
Glucose Level	N
Blood Pressure	N
Skin Thickness(mm)	N
Insulin	N
BMI	N
Age	N
Job Type(Office-work/Field-work/Machine-work)	No
Outcome	C

2. Data Pre-processing

This stage of the model deals with conflicting data in order to produce more accurate and exact findings. There are some missing values in this dataset. Because these attributes cannot have zero values, we imputed missing values for a few selected attributes such as glucose level, blood pressure, skin thickness, BMI, and age. The dataset is then scaled to normalize all values.

3. Clustering

In this phase, we used K-means clustering on the dataset to categorize each patient as diabetic or non-diabetic. Before using K-means clustering, highly associated characteristics such as glucose and age were discovered. For these two attributes, K-means clustering was used. Following its implementation,

Using clustering, we obtained class labels (0 or 1) for each of our records.

Algorithm 1:

```
Generate training set and test set randomly.
Specify algorithms that are used in model
mn=[ KNN( ), DTC( ), GaussianNB( ),
LDA(),SVC(),LinearSVC(),AdaBoost(), RandomForestClassifier(), Perceptron(),
ExtraTreeClassifier(), Bagging(),
 LogisticRegression(),
GradientBoostClassifier()]
for(i=0; i<13; i++) do
Model= mn[i];
Model.fit();
Model.predict();
Print (Accuracy(i),confusion_matrix, classification_report);
End
```

Algorithm 2: Diabetes Prediction using pipeline

Step1: Import required libraries.

Step2: Import diabetes dataset.

Step3: Create pipeline for algorithms giving highest accuracy.

Step4: Add theses pipeline to a dictionary where all pipelines will be stored.

Step5: Fit the pipelines in training dataset.

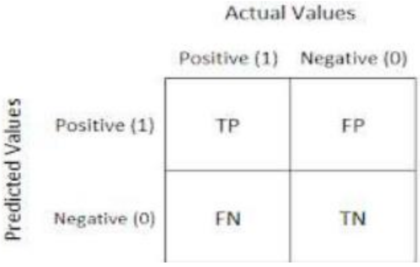
Step6: Compare accuracies of all pipelines added.

Step7: Prediction and identification of the most accurate model will be done on test data. Pipelines work by allowing for a linear sequence of data transforms to be chained together culminating in a modelling process that can be evaluated. The goal is to ensure that all of the steps in the pipeline are constrained to the data available for the evaluation, such as the training dataset or each fold of the cross-validation procedure.

Evaluation: This is the final step of prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score. Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as-

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions Made}}$$

Confusion Matrix- It gives us a matrix as output and describes the complete performance of the model



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Where,

TP: True Positive

FP: False Positive

FN: False Negative

TN: True Negative

Accuracy for the matrix can be calculated by taking average as:

$$Accuracy = \frac{TP+FN}{N}$$

Where, N: Total number of samples

F1 score-It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

$$F1 = 2 * \frac{1}{(\frac{1}{precision}) + (\frac{1}{recall})}$$

F1 Score tries to find the balance between precision and recall. Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier. It is expressed as-

$$Precision = \frac{TP}{(TP + FP)}$$

Recall: It is the number of correct positive results divided by the number of all relevant samples. In mathematical form it is given as-

$$Precision = \frac{TP}{(TP + FN)}$$

IV. RESULTS AND DISCUSSION

After applying various Machine Learning Algorithms on dataset, we got accuracies as mentioned below. Logistic Regression gives highest accuracy of 96%.

Table 2: Accuracy Table

Algorithms	Accuracy
Decision Tree	86%
Gaussian NB	93%
LDA	94%
SVC	60%
Random Forest	91%
Extra Trees	91%
AdaBoost	93%
Perceptron	76%
Logistic Regression	96%
Gradient Boost Classifier	93%
Bagging	90%
KNN	90%

Confusion Matrix for Logistic Regression is given below:

Table 3: Confusion Matrix for Logistic Regression

	Diabetic	Non-diabetic
Diabetic	93	5
Non-Diabetic	4	138

The different performance measures that are being compared are Accuracy, F1-Score, Precision and Recall. The Confusion matrix for the algorithm with highest accuracy is mentioned in Table 2. Visualization of these accuracies helps us to understand variations among them clearly.

Table 4: Comparison between accuracies of PIMA Diabetes Dataset and Diabetes Dataset used in this paper

Algorithms	Accuracy with PIMA Dataset	Accuracy with Diabetes Dataset used in this paper
Logistic Regression	76%	96%
Gradient Boost Classifier	77%	93%
LDA	77%	94%
AdaBoost Classifier	77%	93%

Extra Tree Classifier	76%	91%
Gaussian NB	67%	93%
Bagging	75%	90%
Random Forest	72%	91%
Decision Tree	74%	86%
Perceptron	67%	76%
SVC	68%	60%
KNN	72%	90%

Table 5: Result of Algorithm 2: Using Pipelining, we got highest accuracy of 97.2% for Logistic Regression.

Algorithms	Accuracy
AdaBoost	98.8%
Gradient Boost Classifier	98.1%
Random Forest Classifier	98.1%
Logistic Regression	97.5%
Extra Trees Classifier	96.3%
Linear Discriminant Analysis	95%

V. CONCLUSION

In this phase, we used K-means clustering to identify each patient as diabetic or non-diabetic. Before using K-means clustering, significantly associated characteristics such as glucose and age were discovered. For these two attributes, K-means clustering was applied. Following the execution of this, We obtained class labels (0 or 1) for each of our records by clustering.

VI. REFERENCES

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.

[2] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[4] Dr Saravana Kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.

[5] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[6] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.

[7] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[8] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes

Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[9] Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[10] B.M. Patil, R.C. Joshi and Durga Toshniwal," Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[11] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm", Journal of Computing, Volume 3, Issue 12, December 2011.

[12] Toshita Sharma and Manan Shah, "A comprehensive view of Machine Learning techniques on Diabetes detection". 4, Article Number: 30 (2021).