Diabetes Prediction using Logistic Regression

Abhishek Pal
Computer Science & Technology
Dr. B. C Roy Polytechnic
Durgapur, West Bengal
abhishek.pal@bcrec.ac.in

Sarada Mallik
Computer Science & Technology
Dr. B. C Roy Polytechnic
Durgapur, West Bengal
sarada.mallik@bcrec.ac.in

Rima Dutta
Computer Science & Technology
Dr. B. C Roy Polytechnic
Durgapur, West Bengal
rima.dutta@bcrec.ac.in

***Abstract*— Diabetes Mellitus is a serious condition affecting a large number of people worldwide. Various factors such as age, obesity, lack of exercise, genetic predisposition, lifestyle choices, poor diet, and high blood pressure contribute to the onset of this disease. Individuals with diabetes are at an increased risk for complications like cardiovascular disease, renal disease, heart attack, vision issues, and nerve damage. In hospitals, diabetes diagnosis typically involves conducting numerous tests and providing treatment based on the results. Big Data Analytics has become essential in the healthcare industry, which handles massive volumes of data. By utilizing big data techniques, it is possible to analyze large datasets, uncover hidden patterns, and derive insights to predict outcomes more effectively. However, existing methods for classification and prediction in diabetes diagnosis have limited accuracy. In this paper, we propose an improved diabetes prediction model incorporating additional external factors and standard metrics like glucose levels, BMI, age, and insulin. This new model enhances classification accuracy with an updated dataset and introduces a pipeline framework to further improve prediction accuracy.**

Keywords*— *Diabetes Mellitus; Big Data Analytics; Healthcare; Machine Learning

I. INTRODUCTION

Healthcare sectors manage vast volumes of data, which can be structured, semi-structured, or unstructured. Big data analytics is essential for analyzing these vast datasets, revealing secret findings and trends to extract valuable data. A developing nation such as India, Diabetes Mellitus (DM) has emerged as a particularly severe health issue. Classified as a non-communicable disease (NCD), DM affects a significant portion of the population, with approximately 425 million people worldwide suffering from it as of 2017. Each year, around 2-5 million individuals die due to diabetes-related complications, and it is projected that by 2045, this number will increase to 629 million.[1]

Diabetes Mellitus (DM) is categorized into three types. Type-1, also known as Insulin-Dependent Diabetes Mellitus (IDDM), occurs when the body cannot produce sufficient insulin, requiring patients to receive insulin injections. Type-2, or Non-Insulin-Dependent Diabetes Mellitus (NIDDM), happens when the body's cells fail to utilize insulin effectively. Type-3, or Gestational Diabetes, arises when a pregnant woman experiences elevated blood sugar levels, despite not having previously been diagnosed with diabetes. DM can lead to long-term complications and poses significant health risks.

Predictive Analysis is a technique that employs machine learning algorithms, data mining, and statistical methods to analyze current and historical data, uncover insights, and foretell future events. When applied to healthcare data, anticipated analytics can inform crucial decisions and make accurate predictions, ultimately improving patient care, optimizing resources, and enhancing clinical outcomes. [1] Machine learning, a crucial component of artificial intelligence, allows systems to gain knowledge from previous occurrences without requiring specific programming for each situation. It is essential today, to enable automation and minimizing human error.

The current method for detecting diabetes typically involves laboratory tests, such as fasting blood glucose and oral glucose tolerance tests, which are time-consuming. In this paper, emphasis has been given to developing machine Learning & Data Mining based diabetes prediction models. The paper has been arranged in the following way: Section II is based on a literature review related to previous works on diabetes prediction algorithms. Section III outlines the motivation for this study. Section IV presents the proposed diabetes prediction model. Section V discusses the experimental results, followed by the Conclusion and References.effectively. Type-3, or Gestational Diabetes, arises when a pregnant woman experiences elevated blood sugar levels, despite not having previously been diagnosed with diabetes. DM can lead to long-term complications and poses significant health risks.

Predictive Analysis is a technique that employs machine learning algorithms, data mining, and statistical methods to analyze current and historical data, uncover insights, and forecast future events. When applied to healthcare data, predictive analytics can inform crucial decisions and make accurate predictions, ultimately improving patient care, optimizing resources, and enhancing clinical outcomes. Machine learning, a key aspect of artificial intelligence, enables systems to learn from past experiences without the need for explicit programming for every scenario. It is essential in today's world, enabling automation and minimizing human error.

The current method for detecting diabetes typically involves laboratory tests, such as fasting blood glucose and oral glucose tolerance tests, which are time-consuming. This paper focuses on developing a predictive model using machine learning algorithms and data mining techniques for diabetes prediction. The paper is organized as follows: Section II provides a literature review of past work on diabetes prediction and a taxonomy of machine learning algorithms. Section III outlines the motivation for this study. Section IV presents the proposed diabetes prediction model. Section V

discusses the experimental results, followed by the Conclusion and References.

II. LITERATURE REVIEW

The review of related work highlights various healthcare datasets where different techniques and methods have been applied for analysis and prediction. Numerous predictive models have been developed by researchers using a combination of data mining techniques and machine learning algorithms.

In 2015, **Dr. Saravana Kumar N M, Eswari, Sampath P, and Lavanya S** implemented a system using Hadoop and the MapReduce technique for analyzing diabetic data. This system not only predicts the type of diabetes but also assesses the associated risks. The Hadoop-based system is cost-effective and suitable for healthcare organizations.[4]

Aiswarya Iyer (2015) applied classification techniques to uncover hidden patterns in diabetes datasets. This model used Naïve Bayes and Decision Trees, comparing the performance and effectiveness of both algorithms.[5] **K. Rajesh and V. Sangeetha (2012)** also utilized classification techniques, specifically the C4.5 decision tree algorithm, to identify hidden patterns for efficient classification. In 2008, [8] **Humar Kahramanli and Novruz Allahverdi** combined artificial neural networks (ANN) with fuzzy logic to predict diabetes.[9]

B.M. Patil, R.C. Joshi, and Durga Toshniwal (2010) proposed a hybrid prediction model that first applied the Simple K-means clustering algorithm, followed by a classification algorithm using the C4.5 decision tree . **Mani Butwall and Shraddha Kumar (2015)** introduced a model using the Random Forest classifier to predict diabetes behaviour. [7] **Nawaz Mohamudally and Dost Muhammad (2011)** employed the C4.5 decision tree algorithm, neural networks, K-means clustering, and visualization techniques to predict diabetes.[11]

Figure 1 illustrates the taxonomy of machine learning algorithms that can be employed for diabetes prediction. Selecting the appropriate algorithm involves matching features of the data to existing approaches. Machine learning algorithms are broadly categorized into three types: supervised learning, unsupervised learning, and semi-supervised learning.

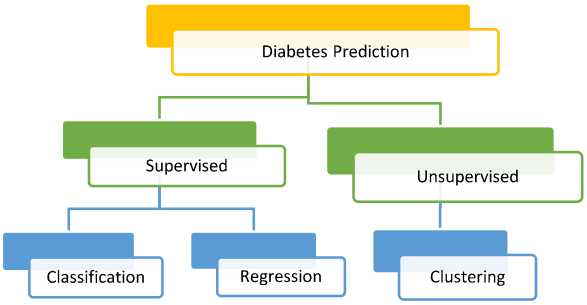


Fig 1: Categorization of Machine Learning for Diabetes Prediction.

A. *Supervised Learning/Predictive Models*
Supervised learning algorithms are employed to create predictive models. These models use known values in a dataset to predict unknown or missing values. Supervised learning works by using a set of input data paired with corresponding outputs, enabling the construction of a model capable of making accurate predictions on new datasets. Common supervised learning techniques include Decision Trees, Bayesian Methods, Artificial Neural Networks, Instance-Based Learning, and Ensemble Methods, which are increasingly popular in the field of machine learning.

B. Unsupervised Learning/Descriptive Models
Unsupervised learning methods are used to develop descriptive models. Unlike supervised learning, unsupervised learning works with known inputs but unknown outputs, making it ideal for analyzing transactional data. Techniques in this category include clustering algorithms, such as k-Means and k-Medians clustering, which are used to group data based on similarities.

C. Semi-Supervised Learning
Semi-supervised learning combines both labeled and unlabeled data to build models. This approach is particularly useful for classification and regression tasks. Examples of regression techniques in semi-supervised learning include Logistic Regression and Linear Regression.

-

III. MOTIVATION

The number of people affected by diabetes has risen significantly over the past decade, largely due to modern lifestyle factors. Current medical diagnostic methods may result in three types of errors:

1. **False-negative**: The test results show a person is not diabetic, even though they are.
2. **False-positive**: The test indicates a person is diabetic when, in reality, they are not.
3. **Unclassifiable cases**: The system fails to diagnose a patient due to insufficient knowledge from past data, leading to an inability to categorize the patient as diabetic or non-diabetic.

These errors in diagnosis can result in treatments that are not necessary or a lack of treatment at the time of need. To minimize these errors and their potential consequences, it is essential to develop a system using machine learning algorithms and data mining techniques that provide accurate results while reducing human intervention.

IV. PROPOSED METHOD

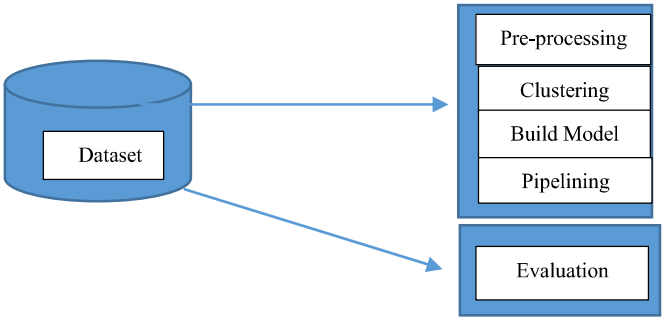


Fig 2: Architectural diagram of predictive model for diabetes.

This architecture consists of five different modules:

i. Dataset Collection

ii. Data Pre-processing

iii. Clustering

iv. Build Model

v. Evaluation

Let's have a look at each model briefly.

i) Dataset Collection

This module involves gathering and analyzing data to identify patterns and trends that aid in prediction and result evaluation. The dataset used in this study is described as follows: it contains over 750 records with 9 attributes.

Table 1. Dataset Information

Attributes	Type
Number of Pregnancies	N
Glucose Level	N
Blood Pressure	N
Skin Thickness(mm)	N
Insulin	N
BMI	N
Age	N
Outcome	C

ii.) Data Pre-processing

This phase addresses inconsistencies in the data to ensure more accurate and reliable results. The dataset includes missing values, therefore we filled in the missing values for essential attributes such as Glucose level, Blood Pressure, Skin Thickness, BMI, and Age, since these attributes cannot be zero. After handling the missing data, the dataset was scaled to normalize all the values.

iii. **Clustering** During this stage, GMM was utilized on the dataset to categorize every patient as either diabetic or non-diabetic. Before conducting GMM, it was discovered that Glucose and Age were highly correlated attributes. These two attributes underwent GMM analysis. Following the application of this clustering method, we obtained class labels (either 0 or 1) for every data point in our dataset.

Algorithm

Initialization:

Randomly initialize the parameters (mean, covariance, and mixing coefficient) for each Gaussian distribution.

The parameters for each Gaussian k include:

Mean: μ_k

Covariance matrix: Σ_k

Mixing coefficient (prior probability): π_k such that

$\sum_{k=1}^{K} \pi_k = 1$ where K is the number of components.

Expectation-Maximization (EM) Algorithm: The EM algorithm is an iterative method used to estimate the parameters of the Gaussian distributions.

Expectation Step (E-step):

Calculate the posterior probability (responsibility) for every data point x_i being generated by each Gaussian component k. This is the probability that data point x_i belongs to Gaussian k, given the current parameters.

The responsibility $\gamma(z_{ik})$is computed as:

$$\gamma_{nk} = \frac{\pi_k N(x_n|\mu_n, \Sigma k)}{\sum_{j=1}^{k} \pi_j \, \mathrm{N}(x_n|\mu_j, \Sigma j)}$$

Where $N(x_n|\mu_n, \Sigma k)$ is the probability density function of the Gaussian distribution.

- **Maximization Step (M-step)**:

Update the parameters based on the posterior probabilities calculated in the E-step:

Update the mixing coefficients: $\pi_k = \frac{N_k}{N}$

Where, $N_k = \sum_{n=1}^{N} \gamma_{nk}$.

Update the mean:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \, x_n$$

Update the covariance matrices:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \, (x_n - \mu_k)(x_n - \mu_n)^T$$

N is the total number of data points, and K is the number of Gaussian components.

- **Convergence**:

Repeat the E-step and M-step until the log-likelihood converges (i.e., changes very little between iterations) or a maximum number of repetitions is reached.

V. Model Building

This is the most important phase which includes model building for the prediction of diabetes. In this, we have implemented Logistic regression algorithms for diabetes prediction.

Logistic Regression Algorithm

1.Initialization:

Define the dataset X (features) and y (target labels), where $y \in \{0,1\}$ (0: no diabetes, 1: diabetes).

Initialize the weights θ (also known as coefficients) and the bias b.

2.Sigmoid Function: Logistic regression predicts the probability P(y=1|x) using the **sigmoid** or **logistic function**. The sigmoid function is defined as:

$h_\theta(x) = \frac{1}{1+e^{-(\theta^t x + b)}}$

Where:

h_θ (x) is the predicted probability.

$\theta^t x$ is the dot product of the weights and input features.

e is Euler's number (the base of the natural logarithm).

3. Cost Function (Log-Loss): The cost function (or loss function) used in logistic regression is the **logistic loss** or **log-loss**, which can be used to minimize the difference between predicted and actual values. The log-loss is defined as:

$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$

where:

m is the number of training examples.

$y^{(i)}$ is the actual label for example *i*.

$h_\theta(x^{(i)})$ is the predicted probability for example *i*.

4. Gradient Descent: Logistic regression uses **gradient descent** to optimize the weights θ and bias b by minimizing the cost function. The gradients of the cost function with respect to the parameters are:

For the weights:

$$\frac{\partial j(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})x_j^{(i)}$$

For the bias:

$$\frac{\partial j(\theta)}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)}).$$

Update the weights and bias using the gradients:

$$\theta_j = \theta_j - \alpha \frac{\partial j(\theta)}{\partial \theta_j}$$

$$b := b - \alpha \frac{\partial j(\theta)}{\partial b}$$

where α is the learning rate, controlling the step size during optimization.

5. Prediction: After training, the model predicts whether a patient has diabetes based on the following rule:

If $h_\theta(x) \geq 0.5$ $predict\ y = 1\ (diabetic)$

If $h_\theta < 0.5$ $predict\ y = 0\ (non\ diabetic)$

Evaluation

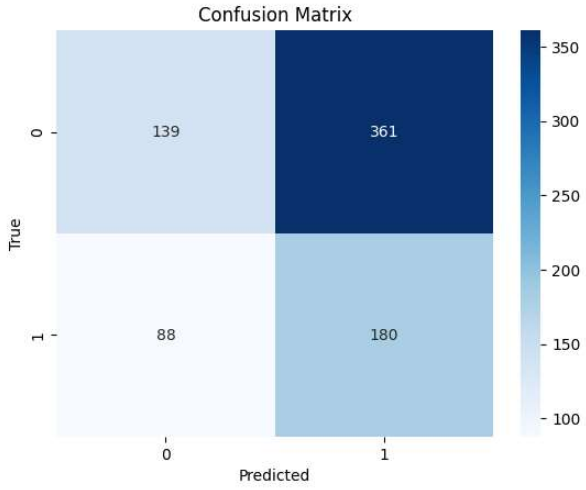
This is the concluding phase of the prediction model, where we assess the prediction outcomes using various evaluation metrics such as classification accuracy, confusion matrix, and f1-score

Classification Accuracy: It represents the proportion of correct predictions out of the total input samples. It can be expressed as:

$$Acuracy = \frac{Number\ of correct\ prediction}{Total\ number\ of\ predictions\ made}$$

- Confusion Matrix-

It provides a matrix as the result and explains the overall model performance.



Where 1st Quadrant FP: False Positive, 2nd Quadrant TP: True Positive,3rd quadrant FN: False Negative, 4th quadrant TN: True Negative.

An average of the values lying across the main diagonal is used to calculate the accuracy of the matrix. It is given as

$Accuracy = \frac{TP+FN}{N}$ where N denotes the total number of samples.

F1 Score tries to find out the balance between precision and recall.

The precision and recall can be calculated as follows

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

VII. RESULTS

After applying Logistic Regression Algorithms on the dataset we got accuracies as 75.32%.

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.80	0.81	99
1	0.65	0.67	0.66	55
accuracy			0.75	154
Micro Average	0.73	0.74	0.73	154
Weighted Average	0.76	0.75	0.75	154

VIII. CONCLUSION

Logistic regression algorithms were utilized in this research to analyze the dataset, with classification performed using the Gaussian Mixture Model (GMM) resulting in an accuracy of 41.54%. Using pipeline logistic regression results in an accuracy of 75.32%. The model enhances the accuracy and precision of diabetes prediction with this dataset in contrast to the current dataset. Moreover, this study could be broadened to evaluate the probability of individuals without diabetes developing the condition in the near future.

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.