Summer Project On

Diabetes Diagnostic Prediction Using ML

By

Vishal Padme(2021510041)

Under the guidance of

Internal Supervisor

Prof. Dr. Pooja Raundale

Department of Master Of Computer Application
Sardar Patel Institute of Technology
Autonomous Institute Affiliated to Mumbai University
2022-23

CERTIFICATE OF APPROVAL

This is to certify that the following students

Vishal Padme(2021510041)

Have satisfactorily carried out work on the project entitled

”Diabetes Diagnostic Prediction Using ML”

Towards the fulfilment of project, as laid down
by
Sardar Patel Institute of Technology
during year
2022-23.

Project Guide:
Prof. Dr. Pooja Raundale

PROJECT APPROVAL CERTIFICATE

This is to certify that the following students

Vishal Padme(2021510041)

Have successfully completed the Project report on

”Diabetes Diagnostic Prediction Using ML”,

which is found to be satisfactory and is approved

at

SARDAR PATEL INSTITUTE OF TECHNOLOGY,
ANDHERI (W), MUMBAI

INTERNAL EXAMINER EXTERNAL EXAMINER

HEAD OF DEPARTMENT PRINCIPAL

Contents

Abstract

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification

1 Introduction

Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Big data analytics is the process which analyses huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data. Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) is classified as Non-Communicable Disease (NCB) and many people are suffering from it. Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to 629 million.[1] Diabetes Mellitus (DM) is classified asType-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person. A technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made. Predictive analytics can be done using machine learning and regression technique. Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes.[1] Machine learning is considered to be one of the most important artificial intelligence features supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case. Machine learning is considered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimum flaws. Existing method for diabetes detection is uses lab tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming. This paper focuses on building predictive model using machine learning algorithms and data mining techniques for diabetes prediction. The paper is organized as followsSection II-gives literature review of the work done on diabetes prediction earlier and taxonomy of machine learning algorithms. Section III-presents motivation behind working on this topic. Section IV gives diabetes prediction proposed model is discussed. Section V gives results of experiment followed by Conclusion and References.

2 Methodoly

The algorithm process proposed in this paper shown in Figure 1. First, the data set as input to the prediction algorithm, and then, though the evaluation model which is the method of introducing a confusion matrix to verify the classification accuracy of the algorithm. Finally, we get the algorithm with the highest accuracy in predicting diabetes.

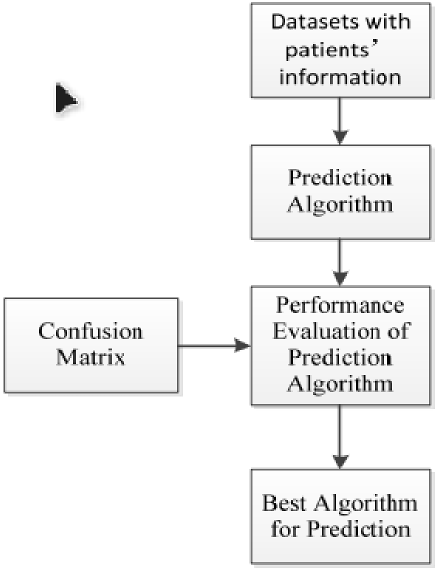

Datasets with patients' information
Prediction Algorithm
Confusion Matrix
Performance Evaluation of Prediction Algorithm
Best Algorithm for Prediction

Fig.1 Process architecture

2.1 Dataset

Used dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
Pima Indian Dataset
Number of Records : 768
Number of Attributes : 9

1. Pregnancies: Number of occurrences of pregnancy

2. Glucose: In a glucose tolerance measure, the plasma glucose concentration after 2 h

3. Blood Pressure The number of times the heart beats per minute is called diastolic blood pressure (mm Hg)

4. Skin Thickness The thickness of the skin folds on the triceps (mm)

5. Insulin: serum insulin (mu U/ml) after 2 h

6. BMI Body mass index

7. Diabetes Pedigree Function Diabetes Pedigree Function

8. Age Age of the person in years

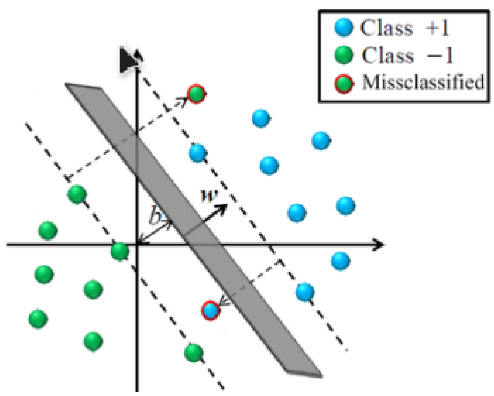
9. Outcome Class variable as a result (0 or 1)

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

3 Algorithms

3.1 Support Vector Machine

SVM is a generalized linear classifier that performs binary classification of data according to supervised learning. Its decision boundary is the maximum-margin hyperplane for solving learning samples [2-4]. SVM uses the hinge loss function to calculate empirical risk and adds a regularization term to the solution system to optimize structural risk. It is a classifier with sparsity and robustness [3]. SVM can perform non-linear classification through the kernel method, which is one of the common kernel learning methods



SVM is an algorithm suitable for binary classification. Zayrit Soumaya [6] and others apply genetic algorithms and SVM to extract features from speech signals to detect some neurological diseases such as Alzheimer's disease, depression and Parkinson's disease. The best accuracy they got was 91.18Agrawal, Dewangan [7] and others used the data of 738 patients for experimental analysis. Combining the SVM with the current discriminant analysis algorithm, the best accuracy rate of is 88.10classification capabilities of support vector machines are excellent, especially when a large number of features are involved.

3.2 Naïve Bayes Classifier

Naive Bayes classifier is a series of simple probability classifiers based on the use of Bayes' theorem under the assumption of strong (naive) independence between features. The classifier model assigns class labels represented by feature values to problem instances, and class labels are taken from a limited set. For the given item to be classified, the probability of each category appearing under the condition of the occurrence of the item is solved, whichever is the largest, and the category to be classified is considered to be. This prediction of the most likely class by probability is suitable for diabetic prediction. The specific classification formulas are shown in (1) to (4). Where represents people who are at risk of diabetes, represents people who are not at risk of diabetes, and X is the data set.

$$P(X|X_p) = \prod_{d=1}^{D} P(x_d|x_p) = P(x_1|x_p)P(x_2|x_p)\ldots P(x_D|x_p) \quad (1)$$

$$P(X|X_n) = \prod_{d=1}^{D} P(x_d|x_n) = P(x_1|x_n)P(x_2|x_n)\ldots P(x_D|x_n) \quad (2)$$

$$P(x_d|x_p) = \frac{Total\,(x_d|x_p)}{Total\ \ x_p} \quad (3)$$

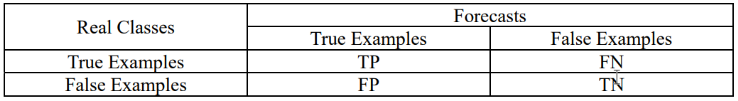
$$P(x_d|x_n) - \frac{Total(x_d|x_n)}{Total\ \ x_n} \quad (4)$$

3.3 LightGBM

LightGBM is a gradient Boosting framework that uses a learning algorithm based on decision trees. It can be said to be distributed and efficient, and has the following advantages: faster training efficiency, low memory usage, higher accuracy, support for parallel learning, and can handle large-scale data. Compared with common machine learning algorithms, its speed is very fast. LightGBM uses histogram algorithm. The basic idea of the histogram algorithm is to discretize the continuous floating-point eigenvalues into k integers, and at the same time construct a histogram with a width of k. When traversing the data, use the discretized value as the index to accumulate statistics in the histogram. After traversing the data once, the histogram accumulates the necessary statistics, and then traverse to find the optimal value according to the discrete value of the histogram.

4 Result

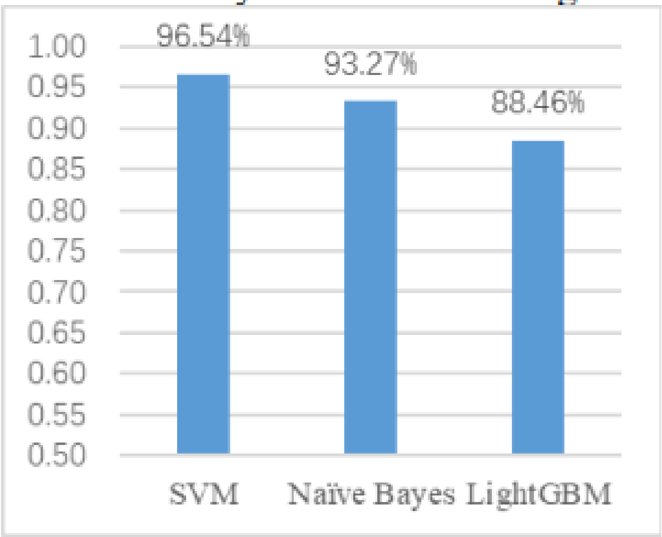
In order to compare the pros and cons of the classification models, it is necessary to provide metrics to evaluate the performance of the models. Here we divide the sample into four classes like true examples (True Positive, TP), false positive (FP), true negative examples (True Negative, TN), and false negative examples (False Negative, FN)[3]. Let TP, FP, TN, and FN respectively denote the corresponding number of samples, TP+FP+TN+FN=n, n is the sample size, and the confusion matrix of the classification result is shown in the following table



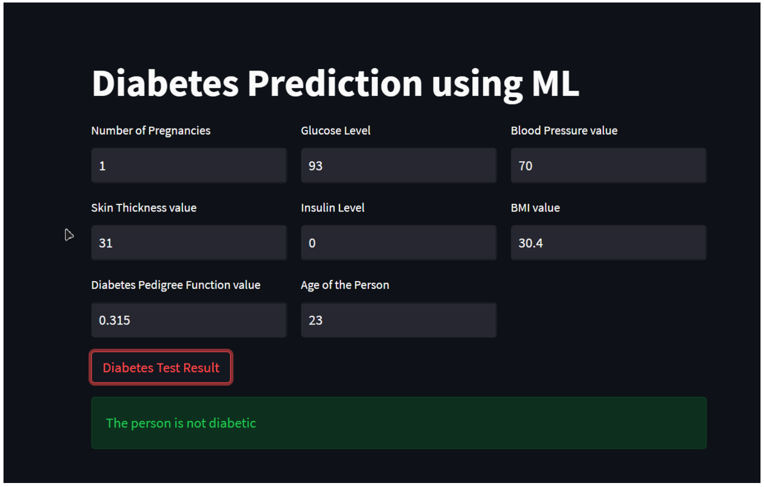
Real Classes	Forecasts	
	True Examples	False Examples
True Examples	TP	FN
False Examples	FP	TN

Here we ,divides the characteristic results into two categories, using ”1” for positive results and ”0” for negative results. First, we split the data into two parts. In this experiment, the ratio of training set to prediction set is 80:20. Using the training set data for model to train, and then use the trained model and prediction set as input in the prediction component. We summarize the results of the above three classification algorithms as shown in Table 2. Although the naive Bayes classifier is the most popular classification algorithm, the final accuracy rate on our data set is only 93.27 percent . SVM has the highest accuracy rate, with an accuracy rate of 96.54 percent . The accuracy of LightGBM is only 88.46 percent . This shows that the most suitable classification algorithm for diabetes prediction is SVM.

Table 4 Accuracy of classification algorithm



SVM Web App Screenshot



Diabetes Prediction using ML

Number of Pregnancies: 1
Glucose Level: 93
Blood Pressure value: 70
Skin Thickness value: 31
Insulin Level: 0
BMI value: 30.4
Diabetes Pedigree Function value: 0.315
Age of the Person: 23

The person is not diabetic

5 Conclusion

Although there is no clear research showing that there is an exact relationship between diabetes and age, there is a clear trend of younger diabetes now. Early detection of diabetes plays a vital role in treatment, and the emergence of machine learning has revolutionized the study of diabetes risk prediction. With the continuous advancement of data mining methods, we have studied various methods of diagnosing diabetes. We found that SVM has the highest accuracy through the confusion matrix evaluation test. However, this kind of research needs to be updated regularly with more instance data sets. Finally, we can see that data mining algorithms through research, machine learning techniques and various other technologies have made outstanding contributions in the medical field and disease diagnosis. It is hoped that it can help clinicians make better judgments on disease status.

6 Bibliography

6.1 Web References

[1.] https://journals.riverpublishers.com/index.php/JICTS/article/download/13397/13153/48699
[2.] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[3.] https://www.sciencedirect.com/science/article/pii/S2666307421000279
[4.] https://www.researchgate.net/publication/339543101_Diabetes_Prediction_using_Machine_Learning_Algorithms
[5.] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[6.] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
[7.] https://www.geeksforgeeks.org/naive-bayes-classifiers/