International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 1717-1721

Research Paper On Diabetes Prediction By Data Mining

AdityaKumarSingh
HarshKulshrestha
RashiBhardwaj
GalgotiasUniversity

Mr. Amit Shukla(ProjectGuide)
CSE Department,GalgotiasUniversity, GreaterNoida (U.P.)

ABSTRACT
Diabetes mellitus is a metabolic disorder characterised via way of means of excessive Blood Sugar. The primary medical sorts are kind 1 diabetes and kind 2 diabetes. Now, the share of younger humans tormented by kind 1 diabetes has extended significantly.Type 1 diabetes is chronic when it occurs in childhood andadolescence, and has a long incubation period. The early symptoms of the onset are not obvious, whichmay lead to failure to detect in time and delay treatment. Long-term high blood sugar can cause chronicdamage anddysfunctionofvarious tissues,especiallyeyes, kidneys,heart, bloodvessels andnerves.
Therefore, the early prediction of diabetes is particularly important. In this paper, we use supervised system-getting to know algorithms like Support Vector Machine (SVM), Naive Bayes classifier and Light GBM to teach at the real facts of 520 diabetic sufferers and capability diabetic sufferers elderly sixteen to ninty five.
Through comparative evaluation of type and reputation accuracy, the overall performance of assist vector system is the best.

Date of Submission: 14-06-2022 Date of acceptance: 28-06-2022

I. INTRODUCTION

Diabetes mellitus is a metabolic ailment with continual hyperglycemia due to a couple of causes. The primary cause is because of defects in insulin secretion and/or function. The traditional signs and symptoms are "3 extra and one less", that is, polyuria, polydipsia, polyphagia and weight loss, which can be observed with the aid of using pores and skin itching. Long-time period carbohydrate, fat, and protein metabolism issues also can motive a whole lot of persistent complications, inclusive of persistent modern disorder, hypofunction, and failure of tissues and organs inclusive of eyes, kidneys, nerves, heart, and blood vessels. Acute and intense metabolic issues can arise in intense situations or beneathneath stress, inclusive of diabetic ketoacidosis (DKA), hypertonic hyperglycemia syndrome. At present, the class standards proposed with the aid of using the WHO Diabetes Expert Committee are:

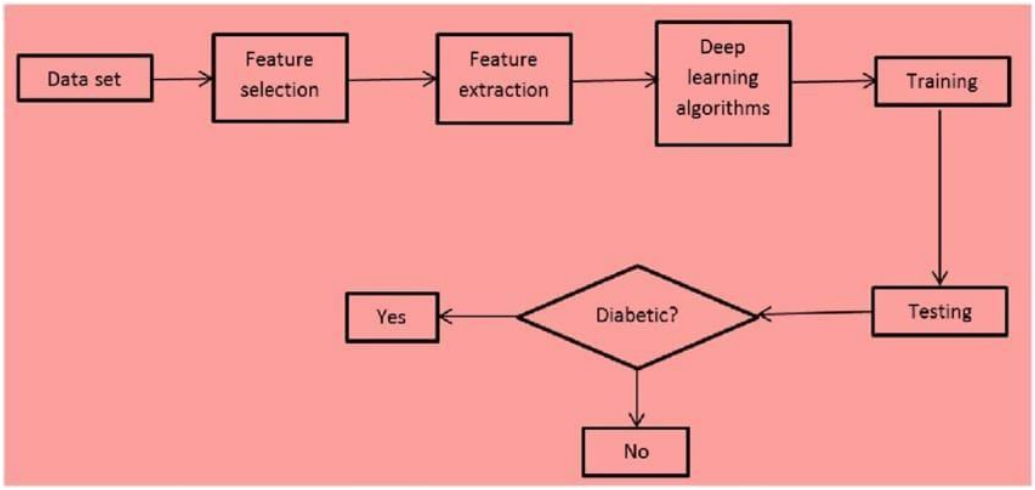
• Type 1 diabetes mellitus (T1DM)
• Immune-mediated (1A): Acute type and slow type.
• Idiopathic (1B): No evidence of autoimmunity.
• Type 2 diabetes mellitus (T2DM):

From insulin resistance with modern insufficient secretion of insulin to the precept plate insulin resistance.

• Special kinds of diabetes, gestational diabetes, etc.

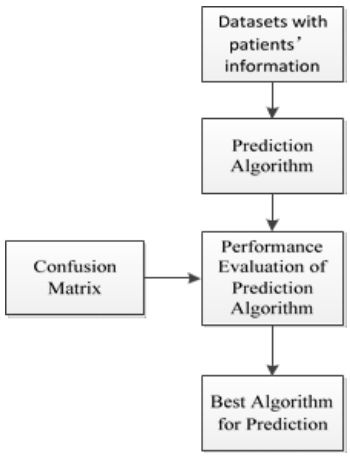
The etiology and pathogenesis of diabetes are fairly complex, and different types have one in all a type causes. Environmental factors play an crucial characteristic withinside the pathogenesis. Environmental factors mainly embody viral infections, chemical poisons and dietary factors. Other symptoms and symptoms and signs that can endorse diabetes embody blurred vision, shortness of breath, chest tightness, slow wound healing, numbness, pores and pores and skin itching, unexpected confusion, coma, periodontal disease, and sexual dysfunction. Common complications embody kidney disease, hectic tool disease, diabetic retinopathy, and macrovascular disease. This may be smooth evidence that, in keeping with WHO, the amount of the diabetic affected individual had been sharply prolonged from 108 million in 1980 to 422 million in 2014[1]. The World Health Organization predicts that with the useful resource of the usage of 2030, diabetes will become the seventh principal purpose of demise withinside the world. The worldwide occurrence of diabetes among adults over 18 years of age prolonged from 4.7% in 1980 to 8.5% in 2014.In the era of big data, and large amounts of data hide various useful information and knowledge. In theprediction of diabetes, a large amount of data filtered

through relevant data sources integrates into adata set for data mining. After that, people can classify and analyze this data set by machine learningalgorithms. This not only allows patients to prevent and treat diabetes at an early stage throughprediction, but also greatly saves time and money costs. This paper uses several algorithms to train theintegrated data set, and finally proposes an appropriate algorithm that can use the early symptoms ofpatientstopredictDiabetes.



II. METHODOLOGY

The set of rules method proposed on this paper proven in Figure 1. First, the records set as enter to the prediction set of rules, and then, aleven though the assessment version that is the approach of introducing a confusion matrix to confirm the class accuracy of the set of rules. Finally, we get the set of rules with the best accuracy in predicting diabetes.



III. Dataset

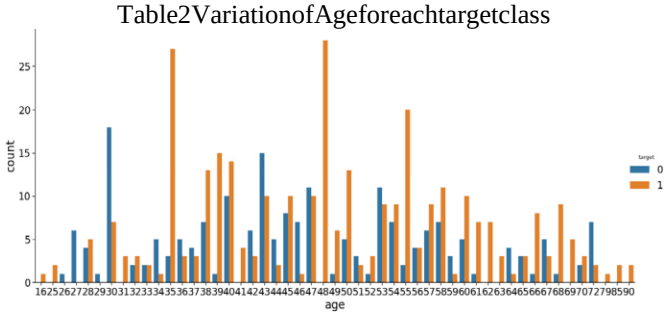
The data set in this article comes from the open-source standard test data set website UCI. The data set was obtained by direct questionnaires from 520 patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh, and was approved by doctors. The data set is divided into 17 attributes including age, gender, polyuria, depression, Sudden weight loss Weakness, Polyphagia, Genital thrush Visual blurring Visual blurring, Keen, Irritability, Delayed healing Partial paresis Muscle stiffness,Alopeciaand Obesity.

Table1Descriptionofattribute

	Attributes	Values
1	Age	16-90
2	Sex	1.Male,0. Female
3	Polyuria	1.Yes,0. No.
4	Polydipsia	1.Yes,0. No.
5	Suddenweightloss	1.Yes,0. No.

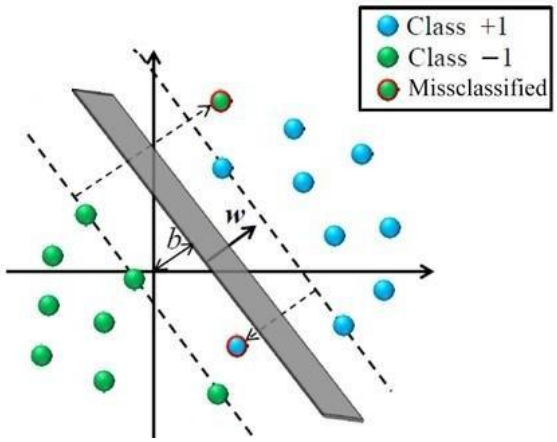
6	Weakness	1.Yes,0. No.
7	Polyphagia	1.Yes,0. No.
8	Genitalthrush	1.Yes,0. No.
9	Visualblurring	1.Yes,0. No.
10	Itching	1.Yes,0. No.
11	Irritability	1.Yes,0. No.
12	Delayedhealing	1.Yes,0. No.
13	Partialparesis	1.Yes,0. No.
14	Musclestiffness	1.Yes,0. No.
15	Alopecia	1.Yes,0. No.
16	Obesity	1.Yes,0. No.
17	Class	1.Positive,0. Negative.

Table2VariationofAgeforeachtargetclass



IV. SupportVectorMachine (SVM)

SVM is a generalized linear classifier that performs binary classification of data according to supervisedlearning. 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 thesolution system to optimize structural risk. It is a classifier with sparsity and robustness [3]. SVM canperform non-linear classification through the kernel method, which is one of the common kernellearningmethods.



SVM is an algorithm suitable for binary classification. Zari Soumya 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 91.18%.

Agrawal, Dewangan 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 88.10%. The classification capabilities of support vector machines are excellent, especially when a large number of features are involved.

NaiveBayesClassifier

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.

LightGBM

Light GBM 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. Light 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.

V. RESULT&DISCUSSION

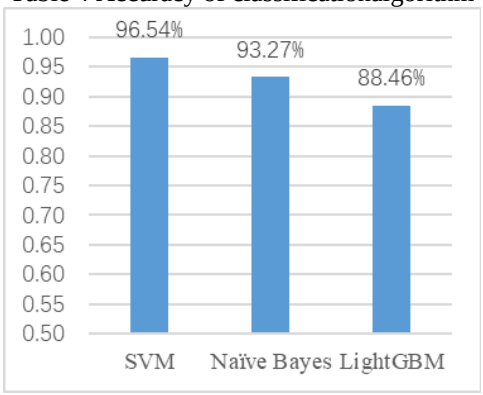
In order to compare the pros and cons of the classification models, it is necessary to provide metrics to evaluate the performance of the models. Here we divide the sample into four classes like true examples (True Positive, TP), false positive (FP), true negative examples (True Negative, TN), and false negative examples (False Negative, FN). 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 3.Table3ConfusionMatrix

RealClasses	TrueExamples Forecasts FalseExamples	
TrueExamples	TP	FN
FalseExamples	FP	TN

This article 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%. SVM has the highest accuracy rate, with an accuracy rate of 96.54%. The accuracy of Light is only 88.46%. This shows that the most suitable classification algorithm for diabetes prediction is SVM.

Table 4 Accuracy of classificationalgorithm



VI. 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.

ACKNOWLEDGMENTS

I would like to express my special thanks of gratitude to the Galgotias University and specially to my guide and reviewer who gave me the golden opportunity to do this wonderful project on the topic "Data mining methods in Diabetes Prediction" which also helped me in doing a lot of Research and I came to know about so many new things. I am really thankful to them.

References

[1]. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression- tutorial-for-machine-learning/#:~:text=What%20is%20a%20Support%20Vector%20Machine%20(SVM
[2].) %3F, -
[3]. So%20what%20exactly&text=This%20is%20exactly%20what%20SVM, on%20the
[4]. %20classes%20to%20predict.
[5]. https://www.researchgate.net/post/How_support_vector_machine_is_used_in
[6]. _prediction https://www.sciencedirect.com/science/article/pii/S1877050920300557