DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS LIKE SVM, NB AND LGBM

Mrs. M. NARMADHA
Assistant Professor
CSE Department
Sridevi women's engineering college
V. n pally, Hyderabad.
Telangana-500075
narmadhamididoddi@gmail.com

Mr. K. SESHAGIRI RAO
Assistant Professor
CSE Department
Narasimha Reddy Engineering College
Miasmaguda,secunderabad.
Telangana-500100
kseshu545@gmail.com

Mrs. K. PRIYANKA
Assistant Professor
CSE Department
Narasimha Reddy Engineering College
Miasmaguda,secunderabad.
Telangana-500100
priyanka.kanuparthi@gmail.com

Abstract

Diabetes mellitus (DM) is a chronic disease that is considered to be life-threatening. It can affect any part of the body over time, resulting in serious complications such as nephropathy, neuropathy, and retinopathy. In this work, several supervised classification algorithms were applied for building different models to predict and classify eight diabetes complications. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. In this paper, we use supervised machine-learning algorithms like Support Vector Machine (SVM), Naive Bayes classifier and Light GBM to train on the actual data of 520 diabetic patients and potential diabetic patients aged 16 to 90. Through comparative analysis of classification and recognition accuracy, the performance of support vector machine is the best.

Keywords: *Diabetes, Machine, Learning, Prediction, Dataset,Support vector machine.*

1. INTRODUCTION:

To reduce the possibility of developing some serious complications related to diabetes, machine learning and data mining techniques can be applied to diabetes-related datasets. Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn. Machine learning itself can be divided into two main categories, namely, supervised and unsupervised learning [6]. The main goal in both cases is to make use of a given dataset to enhance our understanding of the data and discover useful knowledge. Supervised machine learning is characterized by the use of labeled data to train its algorithms and can be utilized for classification or regression tasks. The goal of classification is to assign each unknown instance to one of possible classes or categories for prediction or diagnosis purposes. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to

the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the data can be useful to predict diabetes. Various techniques of Machine Learning can capable to human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected do prediction, however it's tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction. The World Health Organization predicts that by 2030, diabetes will become the seventh leading cause of death in the world. The global prevalence of diabetes among adults over 18 years of age increased from 4.7% in 1980 to 8.5% in 2014[2] In the era of big data, and large amounts of data hide various useful information and knowledge. In the prediction of diabetes, a large amount of data filtered through relevant data sources integrates into a data set for data mining. After that, people can classify and analyze this data set by machine learning algorithms. This not only allows patients to prevent and treat diabetes at an early stage through prediction, but also greatly saves time and money costs. This paper uses several algorithms to train the integrated data set, and finally proposes an appropriate algorithm that can use the early symptoms of patients to predict diabetes.

METHODOLOGY:

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.

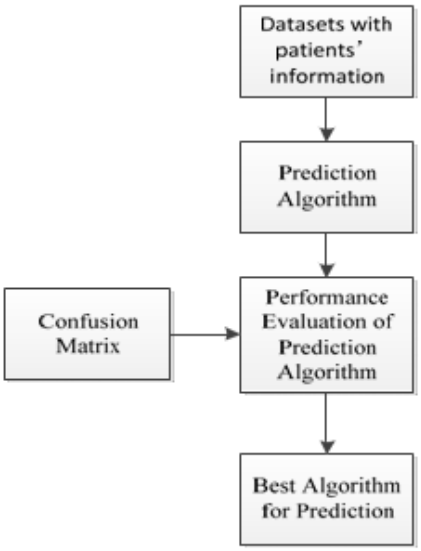


Fig.1 Process architecture

Dataset:

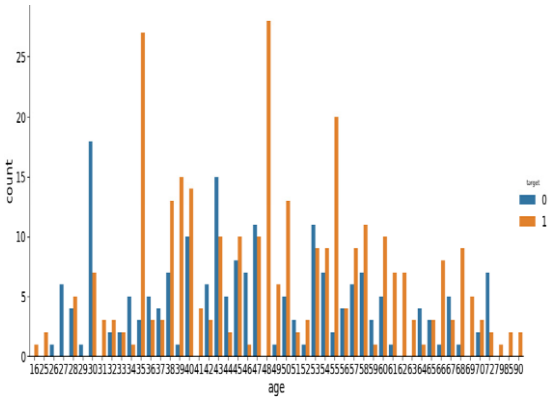
The data set in this article comes from the open source standard test data set website UCI. The data set was obtained by direct questionnaires from 520 patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh, and was approved by doctors. The data set is divided into 16 attributes including age, gender, polyuria, depression, Sudden weight loss 、 Weakness 、 Polyphagia 、 Genital thrush 、 Visual blurring 、 Visual blurring 、 Itching 、 Irritability 、 Delayed healing 、 Partial paresis 、 Muscle stiffness 、 Alopecia 、 and Obesity.

Table 1 Description of attribute

	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	Sudden weight loss	1.Yes, 0.No.
6	Weakness	1.Yes, 0.No.
7	Polyphagia	1.Yes, 0.No.
8	Genital thrush	1.Yes, 0.No.
9	Visual blurring	1.Yes, 0.No.
10	Itching	1.Yes, 0.No.
11	Irritability	1.Yes, 0.No.
12	Delayed healing	1.Yes, 0.No.
13	Partial paresis	1.Yes, 0.No.
14	Muscle stiffness	1.Yes, 0.No.
15	Alopecia	1.Yes, 0.No.
16	Obesity	1.Yes, 0.No.

In Table 1, "1" is to indicate "diseased" and "positive", and "0" is to indicate "not diseased" and "negative". The above attributes are distributed by age among all surveyed patients as shown in Table 2.

Table 2 Variation of Age for each target class



Support Vector Machine (SVM):

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 [5] 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 91.18%. Agrawal, 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 88.10%. The classification capabilities of support vector machines are excellent, especially

when a large number of features are involved.

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, x_n represents people who are not at risk of diabetes, and X is the data set.

$$P(X|x_p) = \prod_{a=1}^{D} P(x_a|x_p) = P(x_1|x_p)P(x_2|x_p)\dots P(x_D|x_p) \quad (1)$$

$$P(X|x_n) = \prod_{a=1}^{D} P(x_a|x_n) = P(x_1|x_n)P(x_2|x_n)\dots P(x_D|x_n) \quad (2)$$

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

$$P(x_a|x_n) = \frac{Total(x_a|x_n)}{Total\ \ x_n} \quad (4)$$

Here D is the attribute with D dimension.

LightGBM:

LightGBM is a gradient Boosting framework that uses a learning algorithm based on decision trees. It can be said to 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 x_p 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.

2. RESULT & DISCUSSION

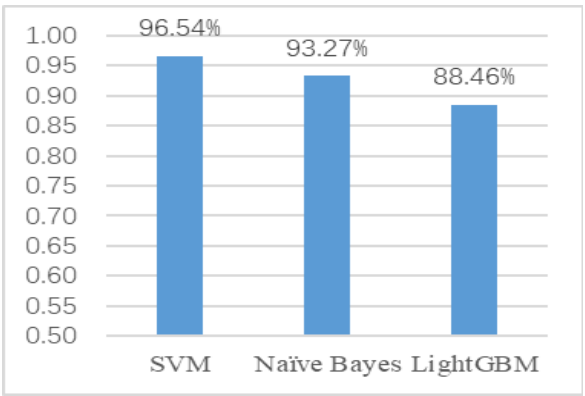
In order to compare the pros and cons of the classification models, it is necessary to provide metrics to evaluate the performance of the models. Here we divide the sample into four classes like true examples (True Positive, TP), false positive (FP), true negative examples (True Negative, TN), and false negative examples (False Negative, FN)[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 3. be distributed and efficient, and has the following advantages: faster training efficiency,

Real Classes	Forecasts True Examples

	False Examples	
True Examples	T P	F N
False Examples	F P	T N

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 LightGBM is only 88.46%. This shows that the most suitable classification algorithm for diabetes prediction is SVM.



3. 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.

REFERENCE:

[1] The genetic algorithm and SVM classifier,Elsevier Ltd Applied Acoustics:2021.doi:10.1016/j

[2] Diabetes, World Health Organization (WHO): 30 Oct 2018.

[3] Vapnik, V.. Statistical learning theory. 1998 (Vol. 3). . New York, NY：Wiley，1998：Chapter 10-11, pp.401-492

[4] Zhou Zhihua. Machine learning. Beijing: Tsinghua University Press， 2016： pp.121-139, 298-300

[5] Li Hang. Statistical learning methods. Beijing: Tsinghua University Press, 2012: Chapter 7， pp.95-135

[6] Qin, J. and He, Z.S., 2005, August. A SVM face recognition method based on Gabor-featured key points. In Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on (Vol. 8, pp. 5144-5149). IEEE. Zayrit Soumayaa, Belhoussine Drissi Taoufiqa, Nsiri Benayadb, Korkmaz Yunusc, Ammoumou Abdelkrim, The detection of Parkinson disease using apacoust.2020.107528

[7] Agrawal, P., Dewangan, A.: A brief survey on the techniques used for the diagnosis of diabetes-mellitus. Int. Res. J. Eng. Technol. (IRJET).02(03) (2015). e-ISSN: 2395-0056; p-

ISSN: 2395-0072

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[9] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[10] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using MachineLearning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.