International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 6, December 2019, pp. 4866~4872
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i6.pp4866-4872

RB-bayes algorithm for the prediction of diabetic in "PIMA Indian dataset"

Rajni[1], Amandeep[2]
[1]School of Computer Application, Lovely Professional University, India
[2]Department of Computer Application, Lovely Professional University, India

Article Info

Article history:

Received Dec 8, 2018
Revised Apr 11, 2019
Accepted Jun 27, 2019

Keywords:

Accuracy
Bayes method
Cross-validation
PIMA Indian data-set
SVM

ABSTRACT

Diabetes is a major concern all over the world. It is increasing at a fast pace. People can avoid diabetes at an early stage without any test. The goal of this paper is to predict the probability of whether the person has a risk of diabetes or not at an early stage. This would lead to having a great impact on their quality of human life. The datasets are Pima Indians diabetes and Cleveland coronary illness and consist of 768 records. Though there are a number of solutions available for information extraction from a huge datasets and to predict the possibility of having diabetes, but the accuracy of their mining process is far from accurate. For achieving highest accuracy, the issue of zero probability which is generally faced by naïve bayes analysis needs to be addressed suitably. The proposed framework RB-Bayes aims to extract the required information with high accuracy that could survive the problem of zero probability and also configure accuracy with other methods like Support Vector Machine, Naive Bayes, and K Nearest Neighbor. We calculated mean to handle missing data and calculated probability for yes (positive) and no (negative). The highest value between yes and no decide the value for the tuple. It is mostly used in text classification. The outcomes on Pima Indian diabetes dataset demonstrate that the proposed methodology enhances the precision as a contrast with other regulated procedures. The accuracy of the proposed methodology large dataset is 72.9%.

Corresponding Author:

Amandeep,
Department of Computer Application,
Lovely Professional University,
Phagwara, Punjab, India.
Email: amandeep.bagga@gmail.com

1. INTRODUCTION

Diabetes is a disease in which blood sugar level is very high. There are two kinds of diabetes. IN Type1 diabetes the body does not deliver insulin. Approximately 10% of all cases are Type1.In Type-2 body does not create enough insulin. Roughly 90% of all instances of diabetes worldwide of this kind [1]. Pima County is a region in the south focal district of the U.S. province of Arizona. PIMA is a standout amongst the most studied population with respect to diabetes, not just among Native Americans yet around the globe [2]. Most of the research on diabetes did use Pima Indian dataset. The dataset contains a record of those diabetic patients that insinuate discrete type 2 positive and negative cases. The course for a diabetes patient to live with this illness is to keep the glucose as typical as great without genuine high or low blood sugars. There are a number of ways through which person can survive with this disease if he/she is using correct routine life. Like, manage your weight, exercising, eating a balanced healthy diet, limiting alcohol intake, controlling blood pressure or using a kind of insulin [3]. According to IDF (International Diabetes Federation), diabetes person is increasing day by day [4]. It has reached 382 million in 2013. When the patient is diabetic and having stress also this leads to microvascular damage [5], unhealthy

cholesterol level also. To make life better, we need to record history about the patient, so that it will help the doctors to take the best decision for patient's health.

The major goal of this paper is to create such an algorithm that creates a group of patterns and naturally settle on choice in light of preparing information [6]. To check validation of the algorithm, we are testing on the real-time dataset. This can be called supervised learning. Training data are known as those data. The preparation information includes a course of action of training cases. A supervised learning computation separates the arrangement data and produces an assembled limit, which can be used for mapping new delineations. A perfect situation will take into consideration the calculation to precisely choose the class marks for the unseen tuple. This requires the taking of computation n to whole up from the preparation information to an unseen tuple in a "sensible" way. Although unsupervised techniques have been used in diagnostic many diseases [7]. One unsupervised technique is clustering [8-12].

Both supervised and unsupervised techniques has emerged as the most useful way to extract relevant information from huge datasets. Though there are number of solutions available for information extraction, but the accuracy of the mining process is far from accurate. For achieving highest accuracy, the issue of zero probability, which is generally faced by Naive Bayes analysis, needs to addresses suitably. The proposed framework aims to extract the required information with high accuracy that could survive the problem of zero probability.

This technique is sure to be an effective alternative to naïve bayes because of overcoming the problem of zero frequency. When RB-Bayes were applied on this PIMA Indian dataset, the results arrived at gave an accuracy of 72.9%.It can be concluded that RB-Bayes algorithm produced more accurate results than the existing classification algorithm.

Rests of sections are sorted out as follows. In Section 2 we speak to the related work. In Section 3 depicting the examination strategy and all techniques joined to the proposed technique are clarified. In Section 4, execution and results are talked about. Section 5 speaking to a conclusion and future work.

2. RELATED WORK

A number of methods have been used for diabetes classification. [13] used discriminant analysis, SVM and 10 fold cross validation for classification and to check accuracy and it achieve 82.05%. [14] used a general regression neural network for diabetes classification [15]. Proposed a method for diabetes classification is genetic programming. To check the accuracy of the model, [16] author test hybrid model on two datasets. One of them is a Pima Indian dataset and the second one is clever land heart disease. The result of accuracy for both of the model is 84.24% and 86.8%. Using Linear discriminant analysis and Neuro-fuzzy system [16] intelligent diagnosis system was developed for classification. The accuracy was 84.61%. Another intelligent method was proposed to classify diabetes that based on [17] Small-World Feed Forward ANN. This method having the highest accuracy i.e. 91.66%.Set of fuzzy rules extracted for classification of diabetes [18]. By using this method the author achieved an accuracy of 84.24%. For diabetes classification author did a comparative study of diabetes. They used [19] Levenberg-Marquardt algorithm and probabilistic NN for diabetes classification. Calisir used [20] Morlet Wavelet Support vector machine along with Linear discriminant analysis for diabetes classification. Their achieved classification accuracy of 89.74%.For classification one of supervised powerful techniques naïve bayes can also be used for prediction of diabetes.Naive bayes already used for prediction of mobile phone [21]. Other powerful data mining algorithms have been used for prediction of hear diseases [22].

3. RESEARCH METHODOLOGY

Concentrating on expectation and classification of diseases, the present study uses RB-Bayes algorithm based on the Bayes method and did a comparison with naive Bayes, SVM, and decision tree. The general system of proposed demonstrate appears in Figure 1. We propose another classification method for diabetes classification. Firstly to handle missing data, we replace values with mean. An explanation of methodologies is given in Figure 1.

3.1. Dataset

The dataset is a gathering of Native Americans living in a zone comprising of what is presently focal and southern Arizona. Indian individuals who are leaving in Pima having different ecologically based medical problems identified with the decrease in their conventional economy and cultivating. They have the most noteworthy pervasiveness of type 2 diabetes on the planet, considerably more than is seen in different U.S. populace. While they don't have a more serious hazard than different clans, the Pima individuals have been the subject of a concentrated investigation of diabetes. There is an aggregate of 768 preparing

occurrences incorporated into this informational index. Each preparation event has 8 highlights and class variable that gives the name to that preparation case as showed up in Figure 2. The class variable consists of two values either 0 or 1 that indicating whether a person is healthy or having diabetes.0 means the person is not diabetic 1 means the person is diabetic.

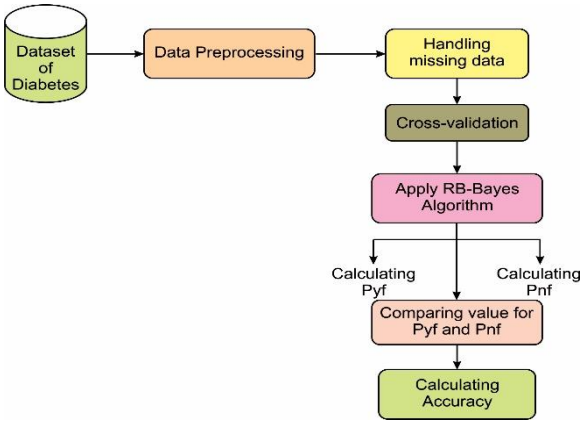

Figure 1. Methodology for prediction

			Min	Max	Average
Pregnancies(a)	Integer	0	0	17	3.845
Glucose(b)	Real	0	44	199	121.682
Blood Pressur...	Real	0	24	122	72.255
Skin Thickness...	Real	0	7	99	26.606
Insulin e	Real	0	14	846	118.660
BMI(f)	Real	0	18.200	67.100	32.451
DiabetesPedig...	Real	0	0.078	2.420	0.472
Age(h)	Integer	0	21	81	33.241

Figure 2. Description of PIMA Indian dataset

3.2. Handling missing data

Handling missing data is an important step before applying the model. If we delete tuple that consists of missing data. This is also a bad idea because might possible that tuple consists of important information. So, to handle missing data, we are having two choices either to calculate the mean or median. In our paper, we are replacing missing values with mean.

3.3. Cross-validations

To avoid the problem of overfitting data, we perform this method on our dataset before finalizing it. Sometimes our model does not perform like the way we expect because we are not reserving part of the dataset on which you do not train the model for testing. Cross-validation is a statistical strategy that in this examination is utilized for execution assessment of learning methodology and execution of a predicted model on an unseen dataset.

Thus, utilizing cross-validation, the informational collections utilized as a part of this exploration are isolated into severally similarly measured subsets. The researcher must ensure that there are sufficient training examples to take in the models. We are using the train_test_split method of class cross-validation in python for dividing the data into preparing and testing part.

3.4. Proposed RB-Bayes algorithm

RB-Bayes is one of simplest supervised technique. It is a classification system in light of Bayes theorem. It is mostly used in text classification. Naive Bayes is also based on the Bayes theorem. But unable to handle the problem of the likelihood of zero possibility. RB-Bayes is proposed to solve this problem [23]. RB-Bayes algorithm provides a way of calculating prediction. Look at Equation (1).

$$P_y F = \frac{T_y}{Total_{Sampleset}} * (\frac{Ty_i + \cdots\cdots\cdots\cdots + Ty_n}{T_F * T_y}) \quad (1)$$

After compairing the value of Pyf and Pnf, preidction can be done whether person is diabetic or not. RB-Bayes classifier has a minimum error rate as compared to other algorithms. All factors are taken into consideration.

4. RESULTS AND COMPARISON OF METHODS

Implementation and consequences of the proposed technique on real-world data sets are clarified in this area.

4.1. Naive Bayes evaluation

In this exploration, Naïve Bayes is connected on the test dataset in excel sheet. Naïve Bayes classifier offers a basic and intense managed characterization technique. The peculiarity of this model is that it expects all information attributes to be of equivalent importance and free of one another. Naive Bayes based on Bayes theorem. Bayes hypothesis can be expressed as Equation (2) [24].

$$P(D|E) = \frac{P(E|D)P(D)}{P(E)} \quad (2)$$

where D and E are events and P (E) ≠ 0.

Apply model applies a model to the real-world dataset. A model is first prepared on an Example Set by another Operator, which is frequently a learning calculation. Subsequently, this model can be connected to another Example Set. Ordinarily, the objective is to get an expectation on concealed information or to change information by applying a pre-processing model. Performance classification operator to check the accuracy of the method. Accuracy with Naïve Bayes is 67.71%.

4.2. Support vector machine

A Support Vector Machine (SVM) is a discriminative classifier formally portrayed by a secluding hyperplane. In a manner of speaking, given named getting ready data (controlled taking in), the computation yields a perfect hyperplane which arranges new outlines. The accuracy of SVM is when applied to a real-world dataset as shown in Table 1.

Table 1. Performance classification using SVM

Accuracy: 70.90%	true 1	true 0
pred. 1	2	0
pred. 0	39	93
class recall	4.88%	100.00%

4.3. Decision tree evaluation

In the choice examination, a choice tree can be used to apparently and explicitly address decisions and fundamental initiative. As the name goes, it uses a tree-like model of choice. A decision tree is applied to the experimental dataset as shown in Figure 3. Set role operator used to define the role of an operator. The part of an Attribute depicts how different Operators handle this Attribute. The default part is consistent, different parts are named special. An Example Set can have numerous special Attributes, yet every extraordinary part can just show up once. In the event that a special role is assigned out to in excess of one Attribute, all parts will be changed to regular aside from the last Attribute. One attribute will be assigned a label attribute. Performance classification operator evaluates the accuracy.

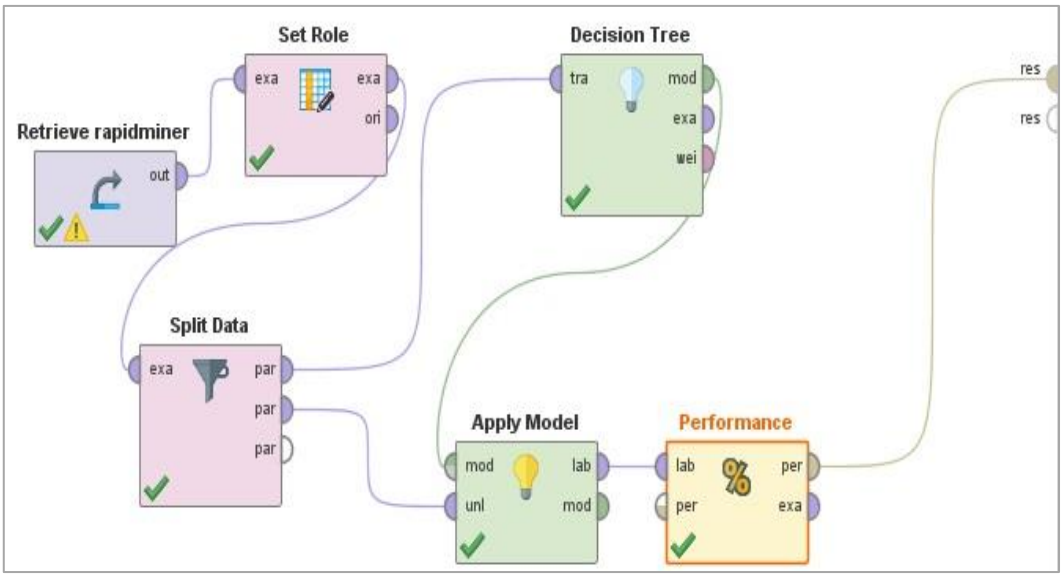

Figure 3. Classifying diabetes using a decision tree

4.4. Performance evaluation of RB-Bayes algorithm

This area gives the test consequences of diabetes infection grouping with proposed RB-Bayes technique in light of Bayes hypothesis. Furthermore, likewise, examination with various strategies also performed and their accuracy is also evaluated. In this study, firstly missing data are handled by calculating mean and missing values are replaced by mean values. Before applying real-world dataset to model, we split the dataset into training and testing. By doing this, we are avoiding the problem of overfitting also. Sometimes we don't get the expected result if we test dataset on training data. Overfitting problem arises and sometimes don't get the expected result. It's better whenever we write any new software, separate the dataset into preparing and testing. Cross-Validation operator is used to utilizing to isolate informational collection into preparing and testing. We calculated a value for the probability of yes (Pyf) and the probability of no (Pnf). Highest value will decide the value for the label. To check precision we tally an incentive for false positive, genuine positive, genuine negative and false negative. The result is shown in Figure 4. This shows the effectiveness of consolidating cross-validation and RB-Bayes calculation for the order precision of diabetes illness. The execution of classifiers that were contrasted and our strategy appears in Table 2.

```
...: c1 = f1.readlines()
...: c2 = f2.readlines()
...: from sklearn.metrics import accuracy_score
...: print('Accuracy score:',accuracy_score(c1, c2))
Accuracy score: 0.7291666666666666
```

Figure 4. Accuracy evaluation of RB-Bayes algorithm

Table 2. Correlation of proposed strategy with different classifiers for Pima Indian

Method	Accuracy
Naive Bayes	67.71%
Support Vector Machine	70.90%
Decision tree	68.18%
RB-Bayes(Proposed method in this study)	72.9%

5. CONCLUSION

In this paper, we propose another procedure in light of Bayes hypothesis for diabetes sickness arrangement utilizing machine learning methods. We calculated mean to handle missing data and calculated probability for yes (positive) and no (negative). The highest value between yes and no decide the value for

the tuple. Keeping in mind the end goal to investigate the adequacy of the proposed strategy and approve the framework, a few examinations directed on Pima Indian informational collection. The dataset was taken from kaggle. The result demonstrated that the proposed RB-Bayes strategy used to acquire good classification accuracy and remove the problem of the possibility of likelihood of zero problems that exist in Naive Bayes. The majority of the methodologies, utilized as a part of this investigation, may likewise be pertinent to other arrangement issues. Subsequently, in our future examination, we intend to assess the proposed technique on extra datasets and specifically on substantial datasets to demonstrate the adequacy of the strategy.

REFERENCES

[1] M. Nilashi, *et al*., "Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset," *Fuzzy Inf. Eng.*, vol. 9, pp. 345-357, 2017.

[2] W. C. Knowler, *et al*., "Diabetes mellitus in the Pima Indians: genetic and evolutionary considerations.," *Am. J. Phys. Anthropol.*, vol. 62, pp. 107-14, 1983.

[3] B. A. Hamburg and G. E. Inoff, "Relationships between behavioral factors and diabetic control in children and adolescents: a camp study," *Psychosom Med*, vol. 44, pp. 321-339, 1982.

[4] V. A. Kumari and R. Chitra, "Classification Of Diabetes Disease Using Support Vector Machine," *Int. J. Eng. Res. Appl. www.ijera.com*, vol. 3, pp. 1797-1801, 2013.

[5] N. H. Barakat, *et al*., "Intelligible support vector machines for diagnosis of diabetes mellitus.," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, pp. 1114-1120, 2010.

[6] R. Bhalla, "Opinion mining framework using proposed RB-Bayes," *Int. J. Electr. Comput. Eng.*, vol. 9, pp. 1-12, 2018.

[7] E. R. Hruschka and N. F. F. Ebecken, "Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach," *Neurocomputing*, vol. 70, pp. 384-397, 2006.

[8] C. H. Chen, "A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection," *Appl. Soft Comput. J.*, vol. 20, pp. 4-14, 2014.

[9] K. Polat, "Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering," *Int. J. Syst. Sci.*, vol. 43, pp. 597-609, 2012.

[10] M. Nilashi, *et al*., "A soft computing approach for diabetes disease classification," *Health Informatics J.*, 2016.

[11] M. Nilashi, *et al*., "Accuracy Improvement for Predicting Parkinson's Disease Progression," *Sci. Rep.*, vol. 6, pp. 1-18, 2016.

[12] M. Nilashi, *et al*., "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telemat. Informatics*, vol. 34, pp. 133-144, 2017.

[13] K. Polat, *et al*., "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," *Expert Syst. Appl.*, vol. 34, pp. 482-487, 2008.

[14] K. Kayaer and T. Yildirim, "Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks," *Iternational Conf. Artif. Neural Networks Neural Inf. Process.*, pp. 181-184, 2003.

[15] M. W. Aslam, *et al*., "Feature generation using genetic programming with comparative partner selection for diabetes classification," *Expert Syst. Appl.*, vol. 40, pp. 5402-5412, 2013.

[16] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Syst. Appl.*, vol. 35, pp. 82-89, 2008.

[17] O. Erkaymaz and M. Ozer, "Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes," *Chaos, Solitons and Fractals*, vol. 83, pp. 178-185, 2016.

[18] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis," *Expert Syst. Appl.*, vol. 38, pp. 14650-14659, 2011.

[19] H. Temurtas, *et al*., "A comparative study on diabetes disease diagnosis using neural networks," *Expert Syst. Appl.*, vol. 36, pp. 8610-8615, 2009.

[20] D. Çalişir and E. Dogantekin, "A new intelligent hepatitis diagnosis system: PCA-LSSVM," *Expert Syst. Appl.*, vol. 38, pp. 10705-10708, 2011.

[21] R. Bhalla and A. Amandeep, "A Comparative Analysis of Factor Effecting the Buying Judgement of Smart Phone," *Int. J. Electr. Comput. Eng.*, vol. 8, pp. 3057-3069, 2018.

[22] M. Abdar, *et al*., "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," *Int. J. Electr. Comput. Eng.*, vol. 5, pp. 1569-1576, 2015.

[23] R. Bhalla and A. Bagga, "Opinion mining framework using proposed RB-bayes model for text classication," *Int. J. Electr. Comput. Eng.*, vol. 9, pp. 477-485, 2019.

[24] S. Russell and P. Norvig, "Artificial Intelligence A Modern Approach," 2013.

BIOGRAPHIES OF AUTHORS

Rajni Bhalla is pursuing PHD from lovely professional University. She is working as assistant professor in Lovely Professional university. She is interested in data mining, data analysis, feature extraction, prediction and clustering techniques.

Amandeep Bagga is an Associated Professor in Department of Computer Application at Lovely Professional university, India. She has completed his studies for B.C.A, M.C.A, Ph.D. from LPU.Her research interest include cloud computing,security,network security and cryptography