Reducing the Risk of Diabetes Using Classification Technique in Data Mining

Puneet Jain
Computer Science
Babulal Tarabai Institute of research & Technology
Sagar, India

Priya Sen
Computer Science
Babulal Tarabai Institute of research & Technology
Sagar, India

Abstract*—Data Mining is used for various purposes in many applications like industries, medical, etc. This is used aimed at removing useful data after a huge quantity of data set. Health monitoring is also used in the data mining concept for predict analysis of the diseases. In health monitoring diabetes is the common health problem nowadays, which affects peoples. In this research, Pima Indians Diabetes Dataset, as well as Waikato Environment for Knowledge Analysis toolkit (WEKA), remained used to associate our consequences through consequences of the existing research work. In the previous research, three classification algorithms were used among which Naïve Bayes stand to be the most efficient one. But due to some limitations, Naïve Bayes is replaced by the Random Forest algorithm (RF). A comprehensive review of literature has been introduced which highlights the research operations using various data mining algorithms and tool as well. The assumption displays that the model achieved higher exactitude of prediction than that previous research.

***Keywords*— data mining, diabetes data, chronic disease, WEKA, Random Forest**

I. INTRODUCTION

This is the process of examining a large amount of data mined expending data mining [DM] to demonstrate the knowledge and to demonstrate it in a simple way to understand the data. Information technology has an important role in the implementation of data mining techniques in different areas like banking as well as education. [2]. medical domain DM area may be utilized effectively to predict diseases by using numerous DM techniques. There are 2 chief purposes for DM. Prediction contains certain variables or areas that are in the procedure of unidentified or future values of other variables. This explanation attentions on discovery patterns that identify factors that donate to diabetes, through human development as well as basic concepts of symptoms, prediction of diabetes and DM approaches. [1]. DM technology may be utilized to predict disease, along through human life, but may also reduce it. Through 2035, it will be doubled to 592 million. [5] This paper studies prediction of diabetes expending various DM methods. The most significant as well as general DM methods for predicting diabetes are classification, clustering, prediction, pure bias as well as decision-making. DM methods are utilized for a variety of applications, such as educational domains and banking.

Diabetes is a chronic disease that affects people around the world. This occurs when the body is unable to produce enough insulin. Unique of most significant hormones in the body, the pancreas secretes insulin and is essentially aimed at preserving equal of glucose. Diabetes may be managed through the assistance of insulin injections, healthy diet as well as consistent implementation. Diabetes indications to other diseases like blindness, hypertension, heart disease as well as kidney disease. Type 1 Diabetes: When the pancreas is unable to yield insulin. Insulin is hormone shaped through the pancreas.

Type 1 Diabetes: This ensues while the pancreas is unable to produce insulin, a hormone that produces pancreas. Type 1 diabetes may transpire at some age. This will happen [7]

Type 2 Diabetes: This happens when the level of insulin is not adequate aimed at body wants. Personal heredity, old age, as well as size increase risk of kind 2 diabetes, which usually transpires at the age of 40 years. [3]

Gestational Diabetes This is 3rd most common procedure, mainly due to high blood sugar levels in pregnant women [4]

Presentational Diabetes: The prediction of diabetes before pregnancy, insulin-dependent diabetes shows a significant role in data mining techniques for presentable diabetes [6]. Numerous DM techniques are practically aimed at prediction of diabetes. These are DM methods provided to predict diabetes.

II. LITERATURE SURVEY

Research paper, "Patient prediction in DM techniques" - A study. The prediction of the disease shows a significant role in DM. This paper studies numerous diseases similar forecast of cardiovascular disease, prediction of breast cancer as well as diabetes, using classification, clustering, decision trees and pure bias methods to predict diabetes. This letter discusses predictive as well as descriptive types of forecasts and includes some areas in the information set to predict the values of new variables. Continuously, the description is centered arranged data of patterns interpreted by humans as well as it discusses different algorithms of DM utilized in the arena of paper medical forecasting [1].

The purpose of the research paper, "Diabetes Forecasting as a result of numerous DM organization systems" describes numerous DM classification methods. In this paper, several classification methods are used to predict diabetes [2].

A research paper on new base algorithms aimed at diabetes data set difficulties "explores numerous DM algo approaches to data mining utilized for the prediction of diabetes. In this paper, the most widely used algorithm for diagnosis there is classification and pure bias [3].

The research paper, "Prophecy prediction describes diabetes mellitus technology," perspective of decision, naïve prejudice, nearest neighbor algorithm (KNN), classification as well as clustering. Diabetes can be predicted by using these effective algorithms [4].

Analyzing the analysis of numerous DM techniques aimed at prediction of diabetes focuses on a total number of persons pretentious through diabetes in the world. This paper predicts that the total number of people with diabetes will double with diabetes. This letter aims for the future and discusses 3 categories of diabetes as well as their causes. It similarly usages prediction and classification methods. It delivers accuracy for diagnosis of the disease [5].

The research paper, "Comparative study of the classification of diseases and health care. This letter can be used to discover the greatest classifier after dissimilar classification of this research algorithm disease, classification algo, data mining that predict this disease. The patient describes the data set decision predicts the main objective [6].

The research paper "Overview of Diabetes Expanding DM Techniques" describes a comprehensive overview of existing DM methods used to predict diabetes. It similarly provides type 1, type 2 as well as type 3 diabetes. The goal of diabetes is to help data mining methods such as neighboring algorithms, Bayesian classifier, NV Bayesian classifier, Bayesian. This letter addresses the effect of diabetes on patients with network and all modalities [7].

Chaudhari et al[8] Diagnosis are unique of most significant requests of such a scheme because it is unique of leading reasons of death worldwide. Humans evaluate the input from complex tests in the laboratory and predict the risk of disease created on risk factors for example tobacco smoking, alcohol use, age, family history, diabetes, hypertension, great cholesterol, physical inactivity, and obesity. Are. Investigators usage numerous DM methods toward help condition maintenance professionals diagnosing heart disease. K-Nearest-Neighbors (KNN) is unique of most positive DM techniques utilized in classification difficulties. Recently, researchers have shown that by mixing different classifiers by voting out form another single classifier. This letter explores KNN's application specifically to help health care professionals diagnose heart disease.

Prof. Mythili et al [9] Diabetes mellitus is a metabolic disease, which is recognized only as diabetes, where the level of high blood sugar affects a person. Diabetes is a metabolic disorder, which is not able to produce or use insulin in the right way. This situation occurs once the body doesn't yield sufficient insulin or cells don't reply to insulin formed by them. Blood sugar test is an important way to diagnose diabetes. In addition, many computer approaches have been suggested aimed at diagnosis of diabetes. Altogether these approaches have certain contribution standards, which result in various experiments which are required to be performed in hospitals. This letter suggests a method which aims to reduce no of patients passing through various medical tests, most of whom are considered to be tedious and time-consuming. The recognized parameters are designed to determine diabetes, so that the operator may guess whether he is pretentious by diabetes or not. Backpropagation algo is utilized aimed at diagnosis.

Ahmed et al [10] Heart disease is the main reason for morbidity as well as mortality in modern society. Medical diagnosis is a significant nonetheless complex charge that must be utilized accurately and efficiently to use commanding information investigation tools to abstract valuable data from medical data. Although there is a large body of data available within the health system, effective analytical tools are in charge. Knowledge discovery, as well as DM, have originated many uses in the field of business as well as science. One of the uses is a diagnostic that proves successful in data mining tool. This research paper proposes a diagnosis of cardiovascular disease finished DM, support vector machine (SVM), genetic algorithm, rough set theory, association rules as well as a neural network. In this study, we temporarily researched the following techniques: the notion of judgment as well as SVM is greatest actual for heart disease. Therefore it has been experiential that DM can assistance recognize or predict great or small heart disease.

Thangaraju et al [11] Data mining is a method of large databases already present in order to generate new information. Different types of DM techniques are available. Classification, clustering, association rules, as well as neural networks, are the most significant technologies in DM. In health care manufacturing, DM plays a significant role. DM is utilized in health care industries aimed at the diagnostic procedure. Diabetes is an old condition. This means that it uses paper clustering techniques to compare the comparison of diabetes prediction approaches that have been around for a long time. Now we usage 3 dissimilar kinds of clustering methods named hierarchical clustering; Density created clustering as well as simple K-mean clustering. it is us.

Durairaj et al [12] Neural networks are unique of soft computing technologies may be utilized to predict medical information. The neural network is recognized as a worldwide predictor. Diabetes mellitus or impartial diabetes is a sickness caused by glucose in the blood. Numerous traditional methods are available created on physical as well as chemical tests to diagnose diabetes. Artificial neural networks (ANNs) based systems can be used effectively to estimate the risk of high blood pressure. This improved prototypical divides dataset in 2 groups. Early detection using soft computing techniques can help doctors reduce the disease. The selected data set aimed at classification as well as testing is created on the Pima Indian Diabetes set as of Machine Learning Database (UCI) collection. In this paper, a detailed survey of applications of various soft computing techniques is done for the prediction of diabetes. The objective of the survey is to identify and predict an effective technique.

III. PROPOSED METHODOLOGY

Bayes Theorem:

A Naive Bayes classifier is a potential ML model. The classifier of the classifier is created on Bayes Principle.

$$\boldsymbol{P(A|B) = \frac{P(B|A).P(A)}{P(B)}}$$

By expending Bayes theorem, we may find a possibility that is owing to illusion. Now, B is proof, as well as A, is an assumption. The perception here is that predictions/topographies are free and a special aspect does not move others.

Random Forest:

Random forest is an easy-to-use, ML algo that is created without hyper-parameter tuning, which is frequently the best result. This is most widely utilized algo due to its simplicity as well as may be utilized for classification as well as regression functions. In this post, you are successful to learn how random forest algo works and numerous other important things nearly it

"Simply put: Random forest produces numerous decision views as well as merge them to get more accurate also stable forecasts" (as shown in Figure 1).

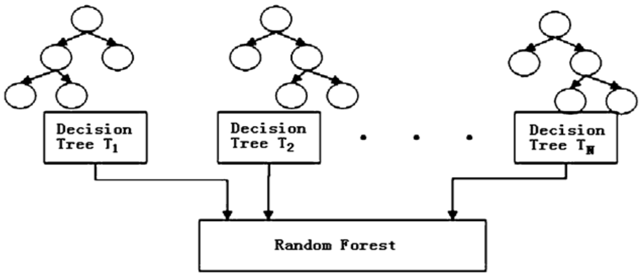


Figure 1 Random Forest

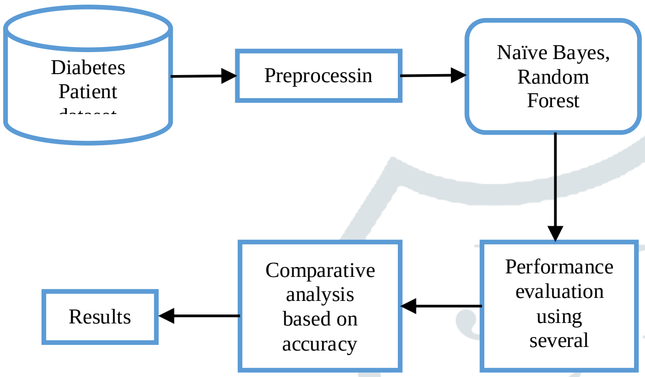

Diabetes Patient dataset
Preprocessin
Naïve Bayes, Random Forest
Performance evaluation using several
Comparative analysis based on accuracy
Results

Figure 2 Proposed Data flow diagram

IV. EXPERIMENTAL RESULTS AND ILLUSTRATIONS

The proposed method has been assessed in Diabetes Dataset (PIDD) [13] which remains taken from UCI collection. This dataset comprises medical details of 768 belongings of female patients. Dataset has similarly comprised 8 properties with a numeric value, negative aimed at negative '0' for a square diabetic and second Class '1' is measured positive for diabetes. Table 1 demonstrations succeeding characteristics that play a main role in the occurrence of heart disease.

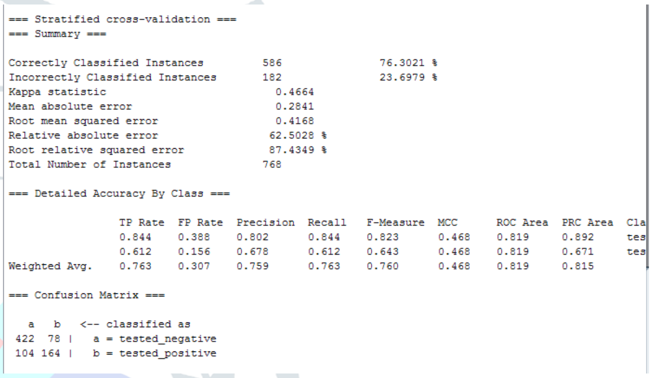
Table 1. Pima Indians diabetes dataset

S. no.	Attribute	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Blood Pressure	Diastolic blood pressure (mm Hg)
4	skin thickness	Triceps skinfold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)
7	Diabetes Pedigree Function	Diabetes pedigree function
8	Age	Age (years)
9	Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0

Now we utilized WEKA 3.9.0 tool to study our datasets. WEKA is a DM scheme established by the University of Waikato in New Zealand that apparatuses DM algo. Basic to refine machine learning (ML) knowledge as well as their applications in a real-world data mining problem.

Now new years, usage of DM algo in medical forecasting study has amplified owing to earnest study in connected parts. Certain investigators have attained significant consequences through this WEKA Toolkit as well as Pema Indian Diabetes Dataset. However, their is a scope aimed at improving accuracy.

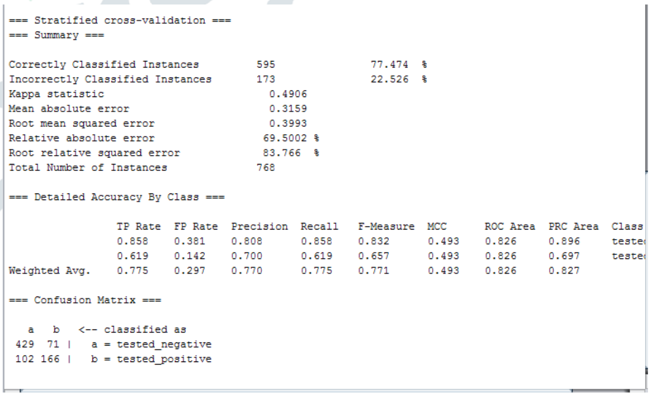
Figure 2 shows the final predicted results of Naïve Bayes algorithm after training the PIDD dataset. The result shows 76.3021% accuracy and the time is taken for the processing of this algorithm is 0.03 seconds.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4664
Mean absolute error                      0.2841
Root mean squared error                  0.4168
Relative absolute error                 62.5028 %
Root relative squared error             87.4349 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.844    0.388    0.802      0.844    0.823      0.468    0.819     0.892     tes
                 0.612    0.156    0.678      0.612    0.643      0.468    0.819     0.671     tes
Weighted Avg.    0.763    0.307    0.759      0.763    0.760      0.468    0.819     0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 422  78 |   a = tested_negative
 104 164 |   b = tested_positive
```

Figure 3 result of Naïve Bayes

Fig 5 displays the accuracy of Random Forest over the existing research models i.e, Naïve Bayes. A comparison has been made which shows that the proposed algorithm has better accuracy than that of the existing algorithm.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         595               77.474  %
Incorrectly Classified Instances       173               22.526  %
Kappa statistic                          0.4906
Mean absolute error                      0.3159
Root mean squared error                  0.3993
Relative absolute error                 69.5002 %
Root relative squared error             83.766  %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.858    0.381    0.808      0.858    0.832      0.493    0.826     0.896     teste
                 0.619    0.142    0.700      0.619    0.657      0.493    0.826     0.697     teste
Weighted Avg.    0.775    0.297    0.770      0.775    0.771      0.493    0.826     0.827

=== Confusion Matrix ===

   a   b   <-- classified as
 429  71 |   a = tested_negative
 102 166 |   b = tested_positive
```

Figure 3 Result of Random Forest

Table 2 Comparison of Training and Simulation Error

Evaluation Criteria	**Classifiers**	
	Naïve Bayes	Random Forest
Kappa Statistics (KS)	0.4664	0.4682
Mean Absolute Error (ABE)	0.2841	0.2266
Root mean squared error (RMSE)	0.4168	0.476
Relative absolute error % (RAE)	62.5028%	49.848%

Root relative squared error % (RRSE)	87.4349%	99.862%

Table 2 displays a comparison of simulation errors as well as the training provided to both the research work using the same data where Random Forest proves to be the better one.

Table 3 Comparing the performance of the students

Algorithm	Correctly classified instances %	Incorrectly classified instances %
Naïve Bayes	76.3021%	23.6979%
Random forest	77.474%	22.526%

Table 3 shows the values of properly as well as imperfectly classified instances along through time duration it took. This comparison clearly shows that the Random Forest algorithm is able to perform better than Naïve Bayes with a short duration of time.

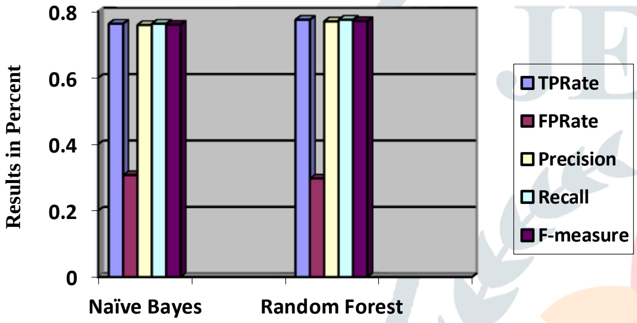


Fig 4Comparing accuracy of classifiers in Base and Propose work

Fig 6 shows the classifier accuracy of the base and propose work with SMO method having higher values than that of the Naïve Bayes.

V. CONCLUSION

Detecting diabetes at its early stage is an important problem of the real world. A lot of efforts are made to provide the best solutions for this problem. In the arena of DM, a number of algo's have been introduced to solve this chronic disease. In this paper, RF classifier is used to predict those patients who are suffering from diabetes in their early sage. In the existing research, the Naïve Bayes classification algorithm is used which RF has replaced in the new research. RF has proven better than NB in standings of accuracy as well as period is taken to perform on the dataset.

In the future, we can use some other technique of data mining to gain more powerful results. We can also perform this operation on any other tool rather than WEKA for the experiment purpose.

REFERENCES

[1] S.Vijiyarani S.Sudha,“ Disease Prediction in Data Mining Technique”– A Survey, International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)

[2] P. Radha, Dr. B. Srinivasan, “Predicting Diabetes by sequencing the various Data Mining Classification Techniques”, IJISET - International Journal of Innovative Science, Engineering & Technology Vol. 1 Issue 6, August 2014

[3] Nilesh Jagdish Vispute, Dinesh Kumar Sahu, Anil Rajput,” a survey on naive Bayes Algorithm for Diabetes Data Set Problems”, International journal for research in Applied Science & Engineering Technology (IJRASET), Volume 3 issue XII, December 2015

[4] Haldurai Lingaraj, Rajmohan Devadass, Vidya Gopi, Kaliraj Palanisamy,” PREDICTION OF DIABETES MELLITUS USING DATA MINING TECHNIQUES”: A REVIEW, Journal of Bioinformatics &Cheminformatics, February 19, 2015.

[5] Dr.M.Renuka Devi, J.Maria Shyla,” Analysis of various Data Mining Techniques to Predict Diabetes Mellitus”, International Journal of Applied Engineering Research ISSN 0973-4562 Vol 11, Number 1(2016).

[6] [6] Isha Vashi, Prof. Shailendra Mishra,” A Comparative Study of Classification Algorithms for Disease Prediction in Health Care”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.

[7] Vrushali Balpande, Rakhi Wajgi,” Review on Prediction of Diabetes using Data Mining Technique”, International Journal of Research and Scientific Innovation (IJRSI) |Volume IV, Issue IA, January 2017 | ISSN 2321–270

[8] Anand A. Chaudhari, Prof.S.P.Akarte, “ Fuzzy and Data Mining based Disease Prediction using K-NN Algorithm”, International Journal of Innovations in Engineering and Technology, Vol. 3, Issue No. 4, April 2014

[9] Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, “Diagnosis of Diabetes Mellitus based on Risk Factors”, International Journal of Computer Applications, Vol.10, Issue No.4, November.2010

[10] Aqueel Ahmed, Shaikh Abdul Hannan, “ Data Mining Techniques to Find Out Heart Diseases: An Overview”, International Journal of Innovative Technology and Exploring Engineering, Vol. 1, Issue No. 4, September 2012

[11] P. Thangaraju, B.Deepa, T.Karthikeyan, “Comparison of Data mining Techniques for Forecasting Diabetes Mellitus”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue No. 8, August 2014

[12] M. Durairaj, G. Kalaiselvi, “ Prediction Of Diabetes Using Soft Computing Techniques- A Survey”, International Journal of Scientific & Technology Research, Vol. 4, Issue No.3, March 2015