2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

Diabetes Disease Prediction Using Data Mining

Deeraj Shetty
BE Computer Engineering
VIVA Institute of Technology
Mumbai, India
sdeeraj28@gmail.com

Kishor Rit
BE Computer Engineering
VIVA Institute of Technology
Mumbai, India
Kisssh.rit@gmail.com

Sohail Shaikh
BE Computer Engineering
VIVA Institute of Technology
Mumbai, India
sohailshaikh352@gmail.com

Nikita Patil
Asst-Prof (Comp. Engg.)
VIVA Institute of Technology
Mumbai, India
nikita.patil@vivacollege.org

***Abstract*— Data mining is a subfield in the subject of software engineering. It is the methodical procedure of finding examples in huge data sets including techniques at the crossing point of manufactured intelligence, machine learning, insights, and database systems. The goal of the data mining methodology is to think data from a data set and change it into a reasonable structure for further use. Our examination concentrates on this part of Medical conclusion learning design through the gathered data of diabetes and to create smart therapeutic choice emotionally supportive network to help the physicians. The primary target of this examination is to assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, we propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.**

***Index Terms*— Disease, Diabetes, Prediction., Naïve Bayes, KNN.**

I. Introduction

Data mining is the investigation of expansive datasets to separate covered up and beforehand obscure examples, connections and information that are hard to recognize with conventional measurable techniques. The territories where data mining is connected as of late incorporate designing, showcasing, human services and monetary anticipating. Data mining in social insurance is also a rising field of high significance for giving what we can say is high anticipation and a more profound comprehension of restorative data. The amount of accessibility of tremendous measure of patient's data which can be used to extricate valuable information, scientists have been utilizing data mining methods to help medicinal services experts in analysis of ailments.

In the usually higher part of the papers, the diabetes forecast system chips away at a little dataset, however our point is to deal with expansive dataset. The quantity of medicinal test required may influence to execution of system in this way we additionally concentrate on diminishing the therapeutic test. It relies on upon which parameter or quality is taken in the system for foreseeing diabetes. Our expectation system will take a shot at a bigger dataset and number of therapeutic testing test required will overcome. Our system utilizes two calculations which we will apply on the same dataset for anticipating diabetes.

A. Problem Definition

The current systems working on diabetes disease prediction works on a small dataset. The aim of our system is to work on a larger dataset to increase the efficiency of the overall system. The number of medical tests also affects the performance of the system; thus, our aim is to reduce the number of medical tests to increase the efficiency of the system.

II. Literature Review

Following is some of the search which has been reviewed for the proposed system: -

Sadegh et al. [6] have proposed, this system that comes under the category of data mining. The system performs data mining on patterns and correlation to predict the economic events. This system utilizes K-Nearest Neighbor for estimating values that will maintain a strategic distance from financial distress and bankruptcy. In the current review k-Nearest Neighbor characterization technique, have been examined for economic estimating. Lately, after the situation of worldwide financial emergency, the quantity of bankrupt organizations has risen. Since organizations' financial distress is the principal phase of bankruptcy, utilizing financial proportions for anticipating financial distress have pulled in a lot of consideration of the scholastics and also economic and financial institutions.

Mohamed EL Kourdi et al. [3] have proposed this system in which Naive Bayes (NB) which is a factual machine learning algorithm is utilized to order Arabic web documents. This system utilizes K-Nearest Neighbor for estimating values that will maintain a strategic distance from financial distress and bankruptcy. In the current review k-Nearest Neighbor characterization technique, have been examined for economic estimating. Lately, after the situation of worldwide financial emergency, the quantity of bankrupt organizations has risen. Since organizations' financial distress is the principal phase of bankruptcy, utilizing financial proportions for anticipating financial distress have pulled in a lot of consideration of the scholastics and economic and financial institutions.

978-1-5090-3294-5/17/$31.00 ©2017 IEEE

Kevin Beyer et al. [10] have proposed this system that tries to explain what happens when dimensionality increases. While dimensionality builds the separation between the nearest and the most distant point gets to be distinctly irrelevant and in this manner the execution is influenced. This may prompt to wrong forecast. Additionally, increment in measurement ought to be dismissed however much as could be expected. In example acknowledgment, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric strategy utilized for order and relapse. The K-NN algorithm is among the easiest of all machine learning algorithms. Both for order and relapse, it can be valuable to relegate weight to the commitments of the neighbors, so that the closer neighbors contribute more to the normal than the more far off ones.

Marius et al. [5] have proposed this system that implements rather fast generating nearest neighbor and appropriate algorithm configuration. In this system, this system they have built up a framework that choses a fitting algorithm in view of the data bolstered which rather creates the fastest nearest neighbor. This algorithm is selected based on dimension of the data. For some PC vision issues, the most tedious segment comprises of nearest neighbor coordinating in high-dimensional spaces. There are no known correct algorithms for tackling these high-dimensional issues that are speedier than straight pursuit. Rough algorithms are known to furnish expansive speedups with just minor misfortune in exactness, however numerous such algorithms have been distributed with just negligible direction on choosing an algorithm and its parameters for any given issue.

III. Proposed System

Flowchart of the system demonstrates the deliberate working of the Diabetes Disease Prediction System. The admin of the system will ask the patient to give provide the current records. Then the admin of the system will ask the patient his/her details needed for the prediction of diabetes disease.

The admin of the system will then choose one of the two appropriate algorithms available. Thus, after using the system, the prediction will be done whether the patient is diagnosed with diabetes or not. If the patient is found diabetic expert recommendations would be provided to the patient so that he/she can recover from diabetes. Whole report will be provided to the patient in the printed form, as usually provided in the hospital's like a report. This system would be very much useful in the field of healthcare.

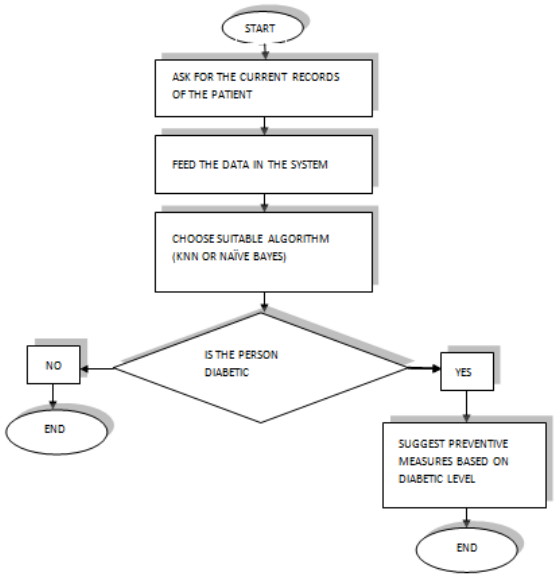

Fig 1. Flowchart

The Execution would be such: -

After coming on to the home page, next the loading and displaying page gets loaded, where the admin of the system must load and display the database of diabetes patients. After this the login page for admin gets loaded, where the admin of the system has login with the unique username and confidential password which is provided to him.

Once the login of the admin has completed successfully, admin must enter the details of the patients such as personal and the factors that is mainly responsible for diabetes [12]. With the help of the details available prediction is performed with the help of Bayesian and K-NN algorithm.

IV. Algorithms

A. Working of naïve bayes:

Steps for Solving Naïve Bayes: -

Step 1: Conversion of training set into a frequency table.

Step 2: Creating what is called Likelihood table that finds the probability which is like Overcast probability = 0.29 and the probability of playing is 0.64.

Step 3: Now, using Naive Bayesian equation, calculate probability for each possible class. The class that has the highest probability among the others is the result of prediction.

TABLE I. NAÏVE BAYES EXAMPLE

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Problem: There will be players playing. Is this statement correct?

This can be solved using posterior probability as discussed below:

TABLE II. FREQUENCY TABLE

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	1	2
Sunny	2	3
Grand total	5	9

TABLE III. LIKELIHOOD TABLE

Likelihood Table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	= 5/14	= 9/14		
	0.36	0.64		

P (Yes in Sunny) = P (of Sunny in Yes) * P (Yes) / P (Sunny)

Now that we possess P (Sunny in Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Also there is P (Yes in Sunny) = 0.33 * 0.64 / 0.36 = 0.60, so this has probability.

Naive Bayes uses almost a similar method that predicts the definite probability of different section based on various attribute values. This algorithm is generally useful in problems that have identical classes.

B. Working of k-nearest neighbor:

Steps For K-NN: -

The example is in below Fig. 2 to acknowledge this algorithm. Following is a wide spread of red circles (RC) and green squares (GS):

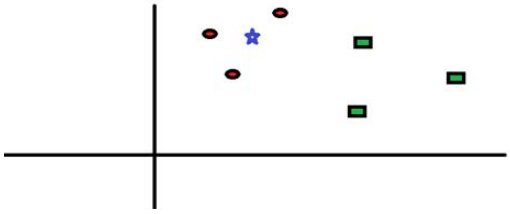

Fig 2. K-NN Example

You expect to discover the class of the blue star (BS). BS can either be RC or GS and that's it. The "K" is KNN calculation is the closest neighbors we wish to take vote from. Suppose K = 3. Consequently, construct a hover with BS as focus similarly as large as to encase just three data points on the same plane. Allude to taking after graph for more points of interest.

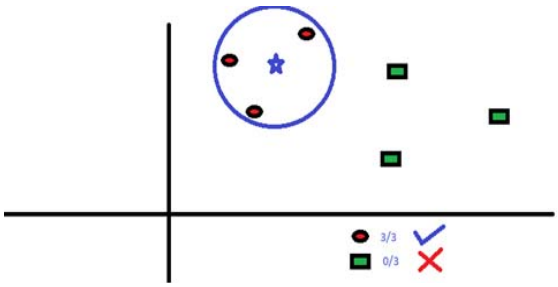

Fig 3. K-NN Example

The three nearest focuses in above Fig 3 to BS is all RC. Subsequently, with great certainty level we can state that the BS ought to have a place with the class RC. Here, the decision turned out to be exceptionally clear as every one of the three votes from the nearest neighbor went to RC.

The decision of the parameter K is exceptionally important in this algorithm. Bayesian model is anything but difficult to

manufacture and especially helpful for substantial data sets. Alongside effortlessness, Naive Bayes is known to beat even profoundly advanced grouping techniques. Bayes theorem provides a simple method of calculating posterior probability P (c in x) from P(c), P(x) and P (x in c).

Look at the equations below:

P (C in X) = P (X in C) P(C)/P(X)

P (C in x) is the posterior probability of class (C, target) given predictor (x, attributes).

P (C) is the prior probability of class.

P (X in C) is the likelihood which is the probability of predictor given class.

P (X) is the prior probability of predictor.

V. Expected result

The goal of our project is to know whether patient is diabetic or not, patient will be diagnosed and it will be depending on the attributes that we are going to take, such as age, pregnancy, pg concentration, tri fold thick, serum ins, body mass index (bmi), dp function, diastolic bp i.e. the factors which are majorly responsible for diabetes.

So, to reduce the correctly know whether the patient is diabetic or not, we are developing a system which will be a prediction system for the diabetes patients. Another best thing about the system is it is will give accurate results whether the patient is diabetic or not with the help of the knowledge base of the larger dataset that we are going to use added the recommendations we are going to provide based on the diabetic levels of the patients. Also, the prediction of the disease will be done with the help of Bayesian algorithm and K-NN algorithm.

VI. Conclusions

By our in-depth analysis of literature survey, we acknowledged that the prediction done earlier did not use a large dataset [12]. A large dataset ensures better prediction. Also what it lacks is recommendation system. When we predict we will give some recommendation to the patient on how to control or prevent diabetes in case of minor signs of diabetes.

The recommendations would be such, that when followed it will help the patient. Thus we will build up a system which will anticipate diabetic patient with the assistance of the Knowledge base which we have of dataset of around 2000 diabetes patients and furthermore to give suggestions on the premise of the nearness of levels of diabetes patients. Prediction will be done with the help of two algorithms Naïve Bayes and K-Nearest Neighbor and also we will compare which algorithm gives better accuracy on the basis of their performance factors. This system which will be developed can be used in HealthCare Industry for Medical Check of diabetes patients.

Future Scope

The proposed system can be developed in many different directions which have vast scope for improvements in the system. These includes:

1. Increase the accuracy of the algorithms.
2. Improvising the algorithms to add more efficiency of the system and enhance its working.
3. Working on some more attributes so to tackle diabetes even more.
4. To make it as a complete healthcare diagnosis system to be used in hospitals.

References

[1] Y. Cai, D. Ji,D. Cai, “A KNN Research Paper Classification Method Based on Shared Nearest Neighbor”, Proceedings of NTCIR-8 Workshop Meeting, 2010.

[2] I. Rish, “An empirical study of the naive Bayes classifier”, T.J. Watson Research Center, 2001.

[3] M.Elkourdi, A.Bensaid, T.Rachidi, “Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm”, Alakhawayn University, 2001.

[4] L.Wang, L.Khan and B.Thuraisingham,“An Effective Evidence Theory based on nearest Neighbor (KNN) classification”, IEEE International Conference, 2008.

[5] M.Muja, David G.Lowe, “Fast Approximate Nearest Neighbors With Automatic Algorithm Configuration”, University of British Columbia.

[6] Sadegh B.Imandoust, M.Bolandraftar, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background”,International Journal of Engineering Research and Applications, Vol. 3, 2013.

[7] Tina R.Patil, S. S. Sherekar, “Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification”, International Journal Of Computer Science And Applications, 2013.

[8] Z.Song, N.Roussopoulos, “*K*-Nearest Neighbor Search for Moving Query Point”,T.J. Watson Research Center.

[9] Harry Zhang, “The Optimality of Naive Bayes”, Faculty of Computer Science at University of New Brunswick.

[10] K Beyer, J Goldstein,R Ramakrishnan and U Shaft, ”When is ‘Nearest neighbor’ Meaningful?” 2014.

[11] Y Cai, D Ji, Dong-feng Cai, “A KNN Research Paper Classification Method Based on Shared Nearest Neighbor”, at Shenyang Institute of Aeronautical Engineering.

[12] K Saxena1, Dr. Z Khan2, S Singh3,” Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm”, at Invertis University.

[13] M Panda and M Patra, “Network Intrusion Detection Using Naïve Bayes” at IJCSNS International Journal of Computer Science and Network Security, VOL7, 2007.

[14] Davis D. Lewis, “Naïve Bayes at Forty – The Independence Assumption in Information retrieval.” AT&T Labs.

[15] Baoli, L., Shiwen, Y. & Qin, L. (2003) "An Improved k Nearest Neighbor Algorithm for Text Categorization, ArXiv Computer Science e-prints.

[16] Chitra, A. & Uma, S. (2010) "An Ensemble Model of Multiple Classifiers for Time Series Prediction", International Journal of Computer Theory and Engineering, 2(3): 1793-8201.

[17] H. Schneiderman and T. Kanade. A statistical method for 3d detection applied to faces and cars. In *Proceedings of CVPR-2000*, 2000.

[18] Milan Kumari, Sunila Godara, "Comparative Study of Data Mining Classification Methods in cardiovascular Disease Prediction", IJCST, Vol. 2, Issue 2, 2011, pp. 304-308