INTELLIGENT DIABETES DISEASE PREDICTION USING DATA MINING AN INNOVATIVE APPROACH TOWARDS HEALTH MONITORING SYSTEM

Dr. Parashuram Baraki[1] , Mr. Jakkappa Godi [2] , Mr. Vinay H Barige[3] , Mr. Nilesh Patil[4]
Ms. Bhagyashri Bisure[5]

[1] Head of Department.(HOD of CSE dept)
[2345] Students of CSE dept
[12345] Computer Science and Engineering.
[12345] HiraSugar Institute of Technology, Nidasoshi, Karnataka, India.

Abstract : *Nowadays, diabetes has become a common disease to the mankind from young to the old persons. The growth of the diabetic patients is increasing day-by-day due to various causes such as bacterial or viral infection, toxic or chemical contents mix with the food, auto immune reaction, obesity, bad diet, change in lifestyles, eating habit, environment pollution, etc. The data analytics is a process of examining and identifying the hidden patterns from large amount of data to draw conclusions. In health care, this analytical process is carried out using machine learning algorithms for analyzing medical data to build the machine learning models to carry out medical diagnoses. This paper presents a diabetes prediction system to diagnosis diabetes. Moreover, this paper explores the approaches to improve the accuracy in diabetes prediction.*

Key Words: Diabetes, Prediction, Naive Bayes, KNN, Decision tree, Knowledge Discovery

I. *INTRODUCTION*

Diabetes is the fast growing disease among the people even among the youngsters. Diabetes is caused by the increase level of the sugar (glucose) in the blood. The diabetes can be classified into two categories such as Type1 diabetes and Type 2 diabetes. Type 1 diabetes is an autoimmune disease. In this case, the body destroys the cells that are essential to produce insulin to absorb the sugar to produce energy.

This type can be caused regardless of obesity. The obesity is the increase of body mass index (BMI) than the normal level of BMI of an individual. Type 1 diabetes can occur in childhood or adolescence age. Type 2 diabetes usually affects the adults who are obese. In this type, the body resists observing insulin or fails to produce insulin. Type 2 generally occurs in the middle or aged groups.

Data mining is process of selecting, exploring and modelling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst. KDD process may consists several steps: like data selection, data cleaning, data transformation, pattern searching i.e. data mining, finding presentation, finding interpretation and finding evaluation.

1.1 Literature survey

This section reviews various research works that are related to the proposed work.

Mohammed Abdul Khaleel et al conducted a survey on data mining techniques on medical data for finding locally frequent diseases. The main focus of this survey is to analysis the data mining techniques required for medical data analysis that is especially used to discover locally frequent diseases such as heart lung cancer, ailments, breast cancer using classification and regression tree (CART) algorithm and the decision tree algorithms such as ID3, C4.5

[1]. Chunhui Zhao et al presented a system for Subcutaneous Glucose Concentration prediction. This proposed model can predict the type 1 diabetes mellitus

[2]. Vaishali Aggarwal et al presented a performance analysis of the competitive learning algorithms on Gaussian data for automatic cluster selection and also studied and analysed the performance of these algorithms and randomized results have been analysed on 2-D Gaussian data with the learning rate

parameter kept simple for all algorithms. Algorithms used in their work include clustering algorithm, competitive learning algorithm and frequency sensitive competitive learning algorithm. Supervised learning machine algorithms are used for classification of the Gaussian data

[3]. K. Srinivas et al developed applications of data mining techniques in healthcare and prediction of heart attacks. This research used medical profiles such as age, sex, blood pressure and blood sugar and predicted the likelihood of patients getting a heart and kidney problems

[4]. Darcy A. Davis proposed individual disease risk prediction based on medical history. This paper also predicts each patient's greatest disease risks based on their own medical history data. Dataset are used for medical coding and collaborative assessment and recommendation engine (CARE) information technique

[5]. From this literature, it is observed that the machine learning algorithms place a significant role in knowledge discovery form the databases especially in medical diagnosis with the medical data.

2 *Problem Definition*

The current systems working on diabetes disease prediction works on few attributes and a small dataset. The aim of our system is to work on a larger dataset to increase the accuracy of prediction. 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.

3. Workflow of the System

Flowchart of the system demonstrates the deliberate working of the Diabetes Disease Prediction System. The admin of the system will ask the patient his/her details needed for the prediction of diabetes disease. Once the details are feed to system then system uses Naïve Bayes Classifier, K- Nearest Neighbor and Decision Tree algorithms to predict the diabetic condition of the patient.

The trained system load and display the predicted value along with patient details. prediction is performed with the help of Naïve Bayes Classifier, K- Nearest Neighbor and Decision Tree algorithms.

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.

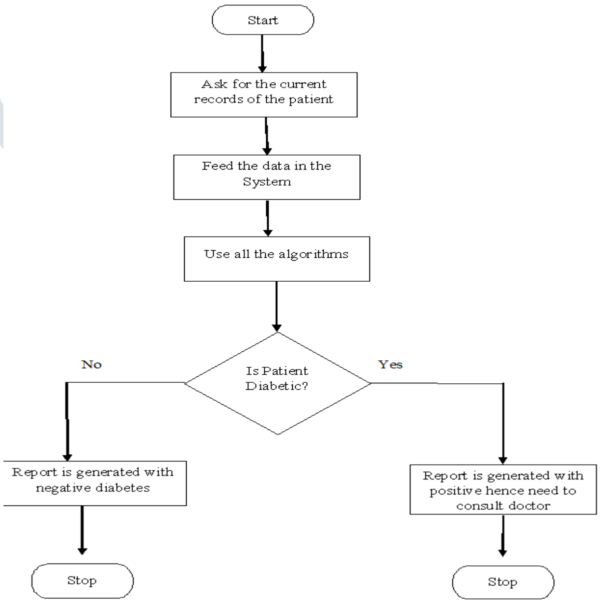

Start
Ask for the current records of the patient
Feed the data in the System
Use all the algorithms
Is Patient Diabetic?
No
Yes
Report is generated with negative diabetes
Report is generated with positive hence need to consult doctor
Stop
Stop

Figure 1. Flow chart

4. Methodology:

In this section the system implementation methodology is explained in detail.

4.1 System Architecture:

In the proposed paper, the data is collected from hospital and it cleaned to avoid redundancy, null attribute values are filled with default values.

The cleaned datasets divided into training dataset and testing dataset. The training dataset feed into Naïve Bayes Classifier, K- Nearest Neighbour and Decision Tree algorithms which create the trained model. With the help of testing dataset the trained model is tested. After successful creation of the model the new patient details, whose diabetes condition need to be predicted are feed into the system.

All three algorithms process input dataset and predict the class of the newly given dataset. The Figure 2 show the system architecture of the proposed system

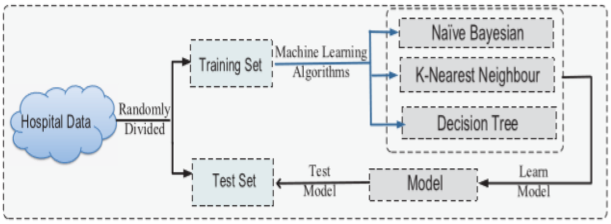


Figure 2. System Architecture

4.2 Working of Algorithms:

In this section the working of the algorithms used for prediction are explained.

4.2.1 Naïve Bayes Algorithm:

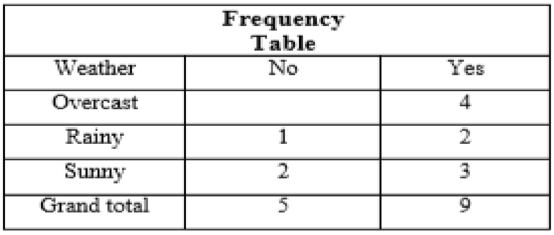
The naïve bayes algorithm works based on the probability theory.Naive bayes algorithm find the posterior probability using prior probability and probability of likelyhood

Steps in Naïve Bayes algorithm: -

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

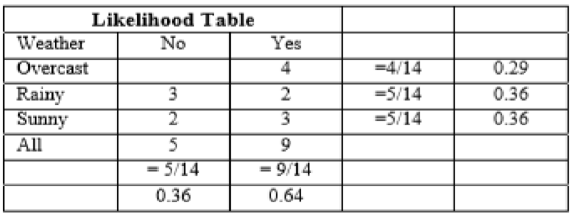
Step 2: Creating Likelihood table that finds the probability which is like Overcast probability = 0.29 and the probability of playing is 0.64.

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



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

Fig 3. Frequency 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		

Fig 4. Likelihood Table

The naïve bayes equation is given below

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

4.2.2 K-Nearest Neighbour :

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Distance functions

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value.

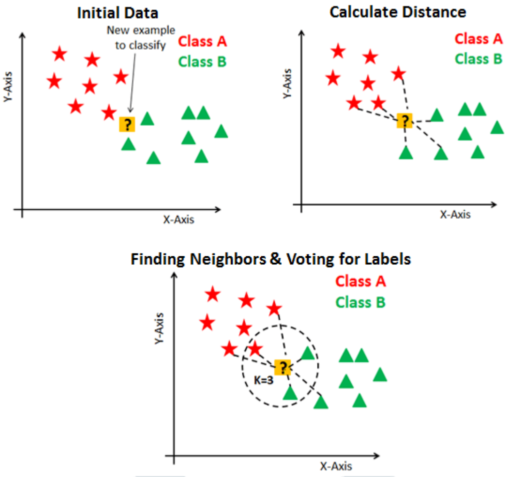


Fig 5. KNN Example

4.2.3 *Decision Tree*:

Decision tree algorithm uses a decision tree to go from observations about an item to conclusions about the item's target value.

1. Place the best attribute of the dataset at the root of the tree.
2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

5.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, gender, skin thickness, glucose concentration, insulin, blood pressure etc.. the factors which are majorly responsible for diabetes. So, to reduce the correctly know whether the patient is diabetic or not, we are developed 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.

6. *CONCLUSION:*

This paper presented a diabetes prediction system for diabetes diagnosis. In order to develop this system, the dataset is collected from the University of California, Irvine (UCI) repository. Different machine learning algorithm namely probabilistic-based naïve Bayes (NB),K-Nearest neighbor (KNN) and decision tree (DT) are used to build the machine learning model to carry out the prediction . Furthermore, the machine learning model is tested with different testing methods such as 10-fold cross validation (FCV), use percentage split with 66% (PS), and use training dataset (UTD) to evaluate the performance of the machine learning model in terms of accuracy. The pre-processing technique is used to increase the accuracy of the model. From the results, it is observed that the pre-processing technique increases the accuracy of the machine learning algorithm except two cases. The pre-processing technique produces better average accuracy for NB compared to other machine learning algorithm.

6. *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] Tina R.Patil, S. S. Sherekar, “Performance Analysis of Naïve Bayes International Journal Of Computer Science And Applications, 2013.

[7] M. Chen, S. Mao, and Y. Liu,
``Bigdata: A survey,"*Mobile Netw.* *Appl.*,vol. 19, no. 2, pp. 171_209, Apr. 2014.

[8] P. B. Jensen, L. J. Jensen, and S.Brunak, ``Mining electronic health records: Towards better research applications and clinical care," *Nature*
Rev. Genet., vol. 13, no. 6, pp. 395_405, 2012.

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