ISSN:0975-9646

Prediction of Diabetes Using Bayesian Network

Mukesh kumari[1], Dr. Rajan Vohra [2],Anshul arora[3]

[1,3]Student of M.Tech (C.E) [2]Head of Department
Department of computer science & engineering
P.D.M College of Engineering
Sector 3A, Sarai Aurangabad, Bhadurgarh

Abstract: This paper helps in predicting diabetes by applying data mining technique. The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in dataset and generate clear and understandable description of patterns. Diabetes mellitus is a chronic disease and a major public health challenge worldwide. Using data mining methods to aid people to predict diabetes has gain major popularity. In this paper , Bayesian Network classifier was proposed to predict the persons whether diabetic or not. The dataset used is collected from a hospital, which collects the information of persons with and without diabetes. We used Weka tool for the experiment and analysis. Classification algorithm is applied on the dataset of persons collected from hospital. Results have been obtained

Keywords: data mining, diabetes ,bayesian network, weka.

INTRODUCTION

Data mining is often described as the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses. Humans, in that sense, are limited by information overload; thus, new tools and techniques are being developed to solve this problem through automation. Data mining uses a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining must also be considered as an iterative process that requires goals and objectives to be specified [1].

Diabetes mellitus (DM) or simply diabetes, is a group of metabolic diseases in which a person has high blood sugar. This high blood sugar produces the symptoms of frequent urination, increased thirst, and increased hunger. Untreated, diabetes can cause many complications. Acute complications include diabetic ketoacidosis and nonketotic hyperosmolar coma. Serious long-term complications include heart disease, kidney failure, and damage to the eyes[2]. There are three main types of diabetes mellitus:

- Type 1 DM results from the body's failure to produce insulin. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes".
- Type 2 DM results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes also with an absolute insulin deficiency. This form was previously referred to as non insulin-dependent diabetes mellitus (NIDDM) or "adult-onset diabetes".
- Gestational diabetes, is the third main form and occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level.

LITERATURE SURVEY

A Research Paper given by **Sudajai Lowanichchai, Saisunee Jabjone, Tidanut Puthasimma** Assistant Professor, Informatic Program Faculty of Science and Technology Nakhon Ratchsima Rajabhat University it proposed the application Information technology of knowledge-based DSS for an analysis diabetes of elder using decision tree. The result showed that the RandomTree model has the highest accuracy in the classification is 99.60 percent when compared with the medical diagnosis that the error MAE is 0.004 and RMSE is 0.0447. The NBTree model has lowest accuracy in the classification is 70.60 percent when compared with the medical diagnosis that the error MAE is 0.3327 and RMSE is 0.454 [3].

In another Research paper presented by **Yang Guo , Guohua Bai , Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden**, The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The dataset used was the Pima Indian diabetes dataset. Pre-processing was used to improve the quality of data. classifier was applied to the modified dataset to construct the Naïve Bayes model. Finally weka was used to do simulation, and the accuracy of the resulting model was 72.3%. [4].

In a Research paper presented by **Ashwinkumar.U.M and Dr Anandakumar.K.R Reva Institute of Technology and Management**, Bangalore S J B Institute of Technology, Bangalore. This Paper has proposed a novel learning algorithm i+Learning as well as i+LRA, which apparently achieves the highest classification accuracy over ID3 algorithm.

The major limitation of our method is the adoption of binary tree rather than multi-branch tree. Such structure increases the tree size, whereas an attribute can be selected as a decision node for more than once in a tree. For that reason, binary trees tend to be less efficient in terms of tree storage requirements and test time requirements, although they are easy to build and interpret[5].

Literature Review on Diabetes, by National Public health :Women tend to be hardest hit by diabetes with 9.6 million women having diabetes. This represents 8.8% of the adult population of women 18 years of age and older in

2003 and a two fold increase from 1995 (4.7%).. By 2050, the projected number of all persons with diabetes will have increased from 17 million to 29 million. [5]

CONCEPTUAL FRAMEWORK

DATA MINING:

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases[7].

BAYESIAN NETWORK:

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis.

One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with bayesian networks offer an efficient and principled approach for avoiding the overfitting of data.

Methods for constructing Bayesian networks from prior knowledge and summarize Bayesian statistical methods for using data to improve these models. With regard to the latter task, we describe methods for learning both the parameters and structure of a Bayesian network, including techniques for learning with incomplete data. In addition, we relate Bayesian-network methods for learning to techniques for supervised and unsupervised learning. We illustrate the graphical-modeling approach using a real-world case study[8].

WEKA TOOL

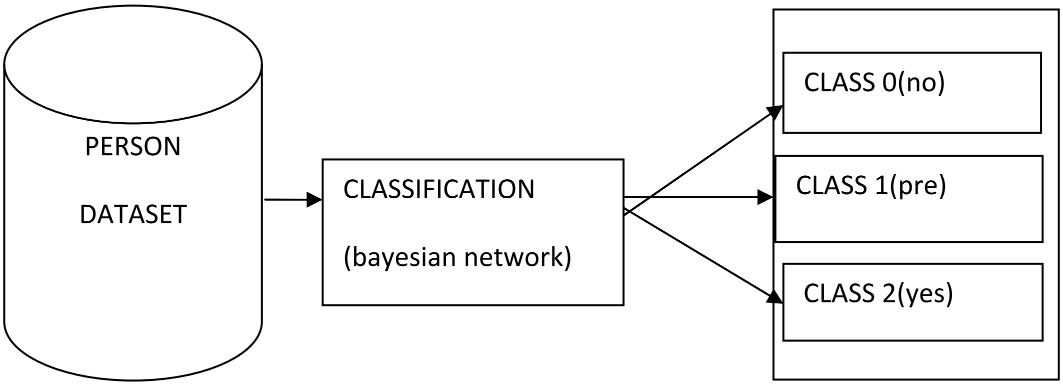
Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code[9].

PROBLEM STATEMENT : Prediction of diabetes using bayesian network

To identify whether a given person in dataset will be diabetic ,non diabetic or pre diabetic will be done on basis of attribute values.Dataset contains all the details of person like fast gtt value, casual gttvalue**,**number of time pregnant,diastolic blood pressure (mmhg),triceps skin fold thickness(mm),serium insulin(µU/ml), body mass index (kg/m)diabetes pedigree function,age of person.

Attributes like fast gtt ,casualgtt,diastolic blood pressure values exceeding a specific value etc may contribute to identify whether a person is diabetic,non diabetic or prediabetic .A brief explanation has been given below through a flow chart .



Flow chart of problem

In this a dataset of person is collected from the hospital and will fed into the software i.e weka which will output the total no of diabetic ,non diabetic persons and pre diabetic persons .The classification will based upon primary attributes value This technique would classify the dataset into three different classes.

RESULTS AND DISCUSSIONS

The dataset that is taken for this research work contains 206 records and 9 attributes for the purpose of predicting whether a person is diabetic or non diabetic based on the symptoms. This dataset is designed in MS excel format.

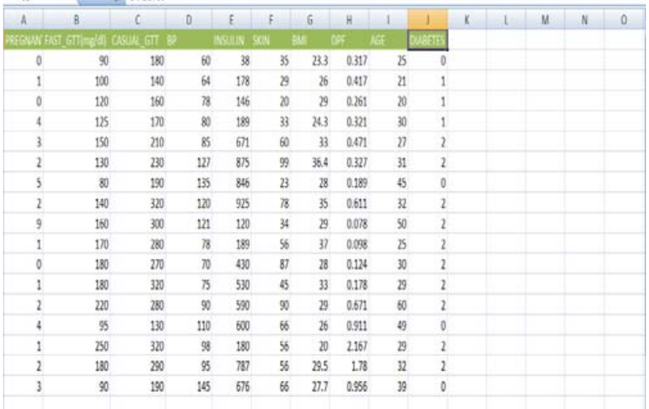
PREPARING DATASET :-

This is the sample of dataset used for prediction.The dataset used contains 206 instances . all instances have 9 input attibutes(X1 to X8) and one output attribute(Y1).table shows the attribute of this dataset .

Attribute no	Attribute	Description	Type
X1	PREGNANT	Number of times pregnant	Numeric
X2	FAST_GTT	Glucose tolerance test	Numeric
X3	CASUAL_GTT	Glucose tolerance test	Numeric
X4	BP	Diastolic blood pressure (mmHg)	Numeric
X5	INSULIN	Serium insulin(µU/ml)	Numeric
X6	SKIN	Triceps skin thickness(mm)	Numeric
X7	BMI	Body mass index(kg/m)	Numeric
X8	DPF	Diabetes pedigree function	Numeric
X9	AGE	Age of person (years)	Numeric
Y	DIABETES	Diabetes diagnose results(no,pre,yes)	Nominal

Attributes of dataset

Sample Dataset[]

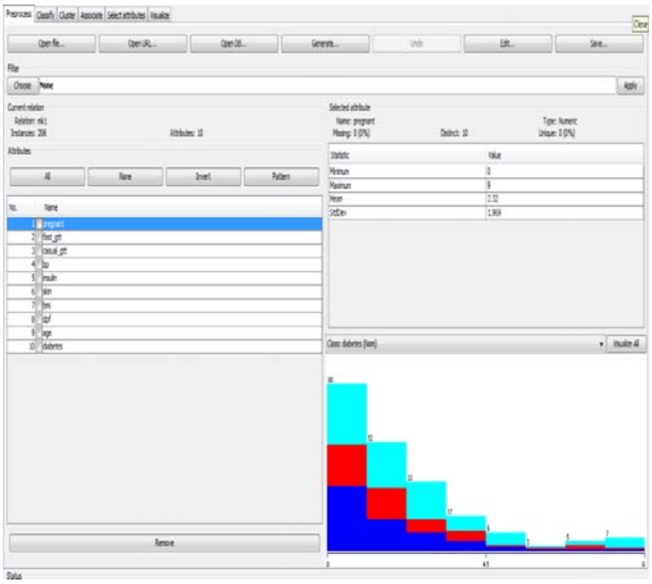


PREGNAN	FAST_GTT(mg/dl)	CASUAL_GTT	BP	INSULIN	SKIN	BMI	DPF	AGE	DIABETES
0	90	180	60	38	35	23.3	0.317	25	0
1	100	140	64	178	29	26	0.417	21	1
0	120	160	78	146	20	29	0.261	20	1
4	125	170	80	189	33	24.3	0.321	30	1
3	150	210	85	671	60	33	0.471	27	2
2	130	230	127	875	99	36.4	0.327	31	2
5	80	190	135	846	23	28	0.189	45	0
2	140	320	120	925	78	35	0.611	32	2
9	160	300	121	120	34	29	0.078	50	2
1	170	280	78	189	56	37	0.098	25	2
0	180	270	70	430	87	28	0.124	30	2
1	180	320	75	530	45	33	0.178	29	2
2	220	280	90	590	90	29	0.671	60	2
4	95	130	110	600	66	26	0.911	49	0
1	250	320	98	180	56	20	2.167	29	2
2	180	290	95	787	56	29.5	1.78	32	2
3	90	190	145	676	66	27.7	0.956	39	0

Figure : Representation of data used for solving problems

Problem :- The 1st problem of work is about predicting whether a person is diabetic or non diabetic in a dataset by applying bayesian network . This problem is solved using the primary attribute . The dataset variables which are used for prediction of diabetes are fast plasma glucose concentration in an oral glucose tolerance test ,casual plasma glucose tolerance test and diastolic blood pressure (mmHg) is decision variable .

If the value of fasting plasma glucose is less than 100mg/dl and value of casual glucose tolerance test is less than 140 mg/dl than it will be given **score 0,** means a person is non diabetic If the value of fasting plasma glucose lies in the range of 100-125 mg/dl and value of casual glucose tolerance test lies in the range of 140-190 mg/dl than it will be given **score 1,** means a person is pre diabetic .If the value of fasting plasma glucose is more than 125mg/dl and value of casual glucose tolerance test is more than 190 mg/dl than it will be given **score 2,** means a person is diabetic



Filter: None

Current relation
Relation: nk1
Instances: 206
Attributes: 10

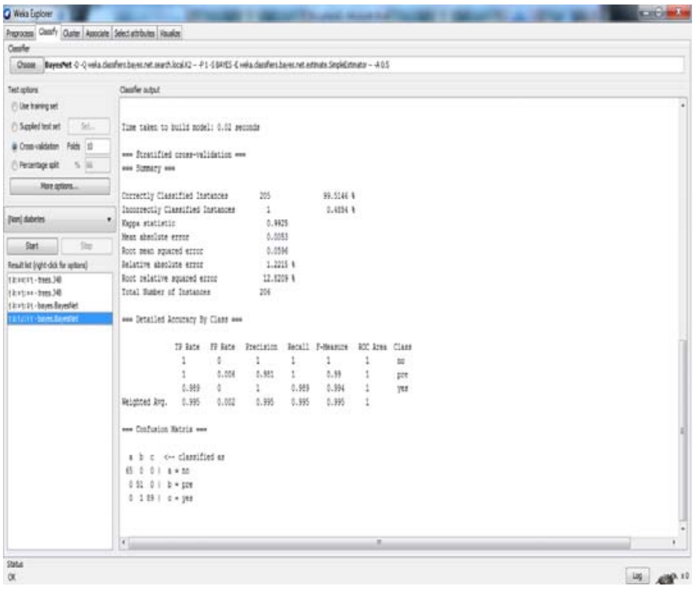
Selected attribute
Name: pregnant
Type: Numeric
Missing: 0 (0%)
Distinct: 10
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	9
Mean	2.32
StdDev	1.969

No.	Name
1	pregnant
2	fast_gtt
3	casual_gtt
4	bp
5	insulin
6	skin
7	bmi
8	dpf
9	age
10	diabetes

Class: diabetes (Nom)

Figure:Representing data load into wek



Classifier: BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Test options: ● Cross-validation Folds 10

Result list: trees.J48; trees.J48; bayes.BayesNet; bayes.BayesNet

```
Time taken to build model: 0.02 seconds

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

Correctly Classified Instances         205               99.5146 %
Incorrectly Classified Instances         1                0.4854 %
Kappa statistic                          0.9925
Mean absolute error                      0.0053
Root mean squared error                  0.0594
Relative absolute error                  1.2215 %
Root relative squared error             12.8209 %
Total Number of Instances              206

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1         1          1        no
                 1         0.006      0.981      1         0.99       1        pre
                 0.989     0          1          0.989     0.994      1        yes
Weighted Avg.    0.995     0.002      0.995      0.995     0.995      1

=== Confusion Matrix ===

  a  b  c   <-- classified as
 65  0  0 |  a = no
  0 51  0 |  b = pre
  0  1 89 |  c = yes
```

Figure :results obtained using Bayesian network

Result for the Problem : Classification of persons in to three classes:

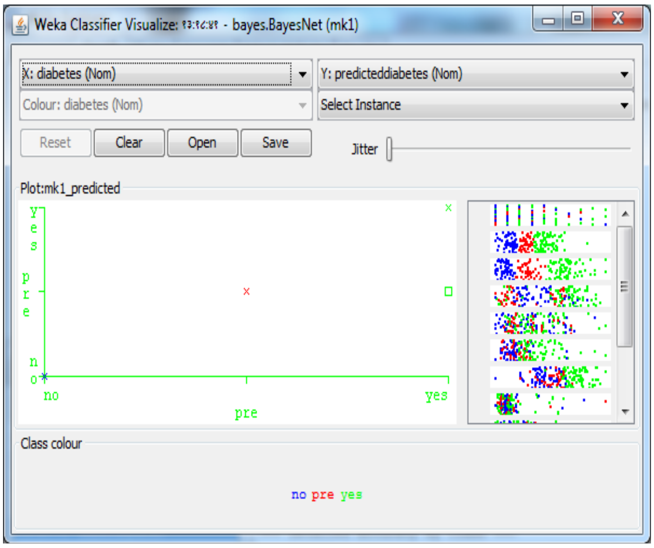
For obtaining the result for the first problem, data of 206 persons is fed to the Weka tool, and as a result the tool generates three classes, i.e., Class no and Class pre and class yes.

Class no: Class predicting person with no diabetes are
Class pre: Class depicting persons with pre diabetes are
Class no: Class predicting person with diabetes are

CLASS	ENTRIES	% OF PERSONS
NO	65	31.55
PRE	51	24.75
YES	90	43.68

Table : Results obtained after applying Bayesian network

In **figure** , the graphical representation of these classes has been shown. The class which is shown using colour 'Blue' is the no class i.e. persons with no diabetes. The class which is shown using 'red' crosses is the pre class i.e those persons which are not diabetic but having symptoms which may lead to diabetes in future. The class which is shown using 'green' crosses is the yes class i.e those persons which are diabetic .



Weka Classifier Visualize: १३:१८:४१ - bayes.BayesNet (mk1)

X: diabetes (Nom)
Y: predicteddiabetes (Nom)
Colour: diabetes (Nom)
Select Instance

Plot:mk1_predicted

Class colour

Figure :Classes shown after applying technique

.

Confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier [9].

	Predicted classes	
Actual classes	TP	FP
	FN	TN
	P	N

Confusion Matrix

- True positive (TP)- These are the positive tuples that were correctly labeled by the classifier [10].If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP)[9].
- True Negative (TN)-These are the negative tuples that were correctly labeled by the classifier [10].
- False Positive (FP)-These are the negative tuples that were incorrectly labeled as positive . However if the actual value is n then it is said to be a false positive (FP) [9].
- False Negative (FN)-These are the positive tuples that were mislabeled as negative [10].

Accuracy is calculated as
(TP+TN)/(P+N)
where, P=TP+FN and N=FP+TN. Or TP+TN/(TOTAL)
According to experimental results, correctly classified instances for bayesian network is 205 Accuracy of bayesian network is 99.51 which is high. Bayesian network is a promising technique for this type of dataset

Conclusion And Future Scope

The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in dataset and generate clear and understandable description of patterns.
This study aims at the discovery of a decision tree model for the diagnosis of diabetes. The dataset used is collected from hospital. Pre-processing is used to improve the quality of data. The techniques of pre-processing applied are attributes identification and selection, data normalization, and numerical discretization.
Next, classifier is applied to the modified dataset to construct the Bayesian model. Finally weka will be used to do simulation, and the accuracy of the model is calculated and compared with other algorithms efficiency.
Classification with Bayesian network shows the best accuracy ,99.51 percent and error in the classification is .48 percent when the results were compared to clinical diagnosis.the mean absolute error (MEA) =.0053 and root mean squared error(MRES =.0596). The total time required to build the model is also a crucial parameter in comparing the classification algorithm.

Future Scope

There are some limitations of this study. Firstly, considering the diabetes dataset, there might be other risk factors that the data collections did not consider. According to , other important factors include gestational diabetes, family history, metabolic syndrome, smoking, inactive lifestyles, certain dietary patterns etc. The proper prediction model would need more data gathering to make it more accurate. This can be achieved by collecting diabetes datasets from multiple sources, generating a model from each dataset.
Secondly, in this study we only use Bayesian network to predict diabetes. Considering of the uncertain factors of some diabetes attributes, in the future work, fuzzy set method will be introduced to improve Bayes Network to do prediction. Also, in order to find a best prediction model, other machine learning methods such as Neural Network will be tested to compare the predicting results.

ACKNOWLEDGEMENTS

Authors would like to thanks to her head Dr. Rajan Vohra, HOD of CSE & I.T department, PDMCE, Bahadurgarh for their valuable support and help.

REFERENCES:

[1] Rohanizadeh.s "A proposed data mining methodogy application to industrial procedures"

[2] en.wikipedia.org/wiki/Diabetes_mellitus

[3] Sudajai Lowanichchai, Saisunee Jabjone, Tidanut Puthasimma, "Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree"

[4] Yang Guo , Guohua Bai , Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, "Using Bayes Network for Prediction of Type-2 Diabetes"

[5] Beckles GLA, Thompson-Reid PE, editors. Diabetes and Women's Health Across the Life Stages: A Public Health Perspective. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion,

[6] Gloria L.A. Beckles and Patricia E. Thompson-Reidy the authors of " Diabetes and Women's Health Across the Life Stages".

[7] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques" Third edition .

[8] en.wikipedia.org/wiki/Bayesian_network

[9]. Sapna Jain 2.M Afshar Aalam3. M. N Doja,"K-MEANS CLUSTERING USING WEKA INTERFACE", Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, February 25 – 26, 2010

[10] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques" Third edition.

[11] Database - "patient data base".