ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

Prediction of diabetes using Classification algorithms

[1] Nandhini.M, [2] Kavitha.R
[1] Student, [2] Assistant Professor
[1] Department Of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India.

Abstract:-- **The main objective of the research is to predict the diabetes patient and Normal patient based on test results or test reports using classification algorithms. In Data mining, different techniques can be used for solving problems. For example, classification, prediction, clustering are data mining techniques. Classification is the process of classify the data according to the features of the data with predefined set of classes. Prediction is Used for predicting the class label for new data. The weka tool is used to develop a classifier for predicting the diabetes patient and normal patient. The Diabetes dataset is used for prediction process. The data set can be divided into two subsets.**

The first one is training set and other one is test set. The training Set contains set of attributes with class labels. The test set contains set of attributes and it doesn't contain class labels. It was predicted by classifier or model. The research takes three algorithms such as Naive Bayes, Multilayer Perceptron and IBK. Each algorithm provides best accuracy for prediction process. The accuracy of the Naive Bayes algorithm is 100%.

Keywords:-- **Diabetes patient, Normal patient, Prediction, Accuracy and Classifiers.**

I. INTRODUCTION

Data mining is the process of extracting or mining hidden knowledge from huge amounts of data. The classifier Predicts the Diabetes patient and Normal patient using symptoms and the test results. In this research, there are three types of classifier or model can be used such as Naive Bayes, Multilayer Perceptron and IBK. Diabetes is the increasing sugar level in blood and it will be identified by various symptoms. The symptoms are increasing urine, drowsiness/thirsty, health illness, leg pain, increasing appetite etc. These symptoms were taken as attributes or features for predicting the diabetes and normal patients.

The attribute PPBS is the class label for this research. PPBS is refers to measurements of sugar level in after food and also consider the attribute FBS.

It is refers to measurements of sugar level in before food. The normal value of FBS is 70-100. The normal value of PPBS is above 140. It will be denoted by mg/dl. So, the classifier has taken PPBS as its class label. If PPBS is more than 140 then patient have diabetes. If it less than 140, then patient is normal. The PPBS is referred as class label a. The FBS is referred as class label b.

II. PROPOSED SYSTEM

The classifiers are developed for predicting the diabetes and normal patient from the dataset. The data set can be divided into two subsets. The subsets are training set and test set. The training set contains set of tuples with class labels that were trained by the classifier. The test set is also contains

Set of tuples with unknown class labels that were test set by the classifier. The test set is used for estimating accuracy of the classifier.

Step1: Load the file/dataset (Diabetes dataset) into weka tool in preprocess step.

Step2: Select the classifier, in that choose algorithms. (Naive Bayes, Multilayer Perceptron and IBK)

Step3: In test options, click percentage split option and also give the percentage for splitting the data set into training set and test set.

Example: Training set = 80% and Test set = 20%

Step4: Click more options button, in that choose output predictions as CSV, then click ok.

Step5: After completing all the above steps, click start. The results are displayed in the screen. It shows the correctly classified instances, incorrectly classified instances, prediction on test set, total number of instances and confusion matrix.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

Architecture of the proposed system:

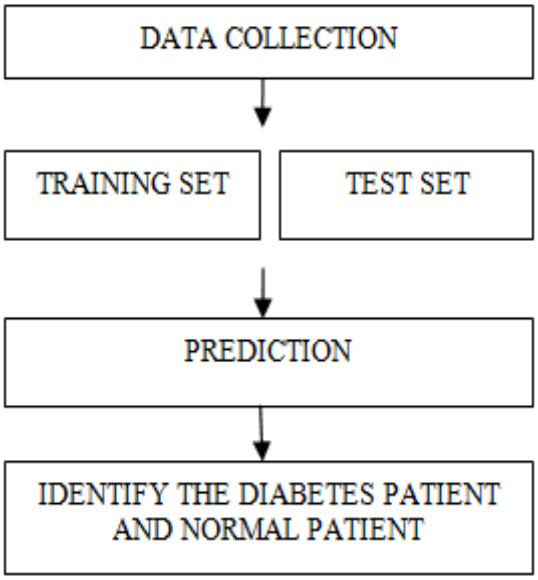

Fig. 1

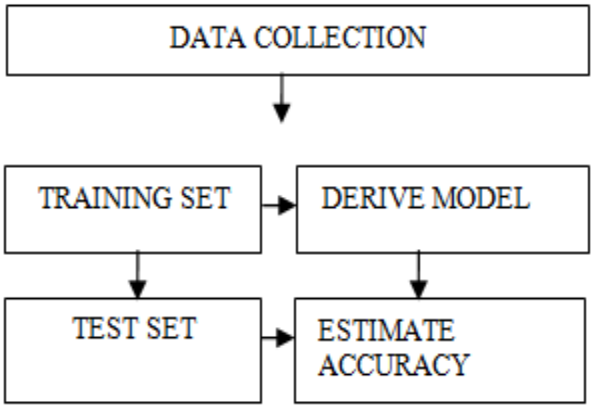

Fig. 2

2.1 Feature Extraction

The first step of the research is to collect the required data for this research. The diabetes dataset can be created. The training dataset contains set of features with predefined class labels. The features or attributes weight, blood pressure, Increasing urine, drowsiness/thirsty, health illness, loss of weight, increasing appetite, eye problem, leg pain, oedema lex, blood sugar(F), blood sugar(PP). Based on this data, the result is produced whether the patient is diabetes patient or normal patient.

2.2 Training Dataset

The training dataset is a set of attributes and its class labels. The class labels is represented by a and b. "a" refers to diabetes patient and "b" refers to normal patient. Each tuple is the information about symptoms and test results of the patient.

The accuracy of a predictor refers to how well a given predictor or classifier can guess the value of the predicted attribute for new or previously unseen data. Here, 80% of the data was trained for prediction.

2.3 Test Dataset

It also contains set of attributes. But, it doesn't contain any class labels. It will be predicted by the classifier. Here, 20% of the data was tested for predicting the class labels. The above process was completed by percentage-split option.

III.UNSUPERVISED LEARNING ALGORITHM

In supervised learning algorithm, the class label of each training tuple is provided. In unsupervised learning algorithm, the class label of each training tuple is not known.

Algorithms

Naïve bayes, Multilayer perceptron, IBK are the classification algorithms. The Prediction was done by using above three data mining algorithms or classifiers. The prediction can be done using these three algorithms with diabetes dataset. Each algorithm gives the output with various parameters.

IV. EXPERIMENT AND RESULTS

The prediction was performed using Naïve bayes, Multilayer perceptron and IBK algorithms on diabetes dataset in weka tool and each algorithm produce the efficient results .These results are called as parameters. The confusion matrix is generated for class label a and b. The class label "a" refers to diabetes patient. The class label "b" refers to normal patient.

Confusion matrix

A confusion matrix illustrates the accuracy of the solution to a classification problem or it is a useful tool for analysing how well your classification can recognize tuples of different classes. Dividing the true-positive by sum of true-positive and false-positive is called as precision. Dividing true-positive by sum of true-positive and false-negative is called as recall. Product of precision and recall divide by sum of precision and recall is called as F-Measures .

True-positive refers to positive tuples that were correctly labelled by the classifier. True-negative refers to negative tuples that were correctly labelled by the classifier.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

False-positive refers to positive tuples that were incorrectly labelled by the classifier. False-negative refers to negative tuples that were incorrectly labelled by the classifier .

Table1. Class Label Prediction

True-Positive	Yes	Yes
True-Negative	No	No
False-Positive	Yes	No
False-Negative	No	Yes

Description:

True-positive refers to class label "a" will be correctly labelled by the classifier.
True-negative refers to class label "b" will be correctly labelled by the classifier.
False-positive refers to class label "a" will be incorrectly labelled by the classifier.
False-negative refers to class label "b" will be incorrectly labelled by the classifier.

Results of classification algorithms:

Table2. Predictions on Test Split

Instances	Actual	Predicted	Error	Prediction
1	2:b	2:b		0.999
2	1:a	1:a		0.989
3	1:a	1:a		0.998
4	2:b	2:b		0.994
5	2:b	1:a	+	0.644

The first tuple was correctly predicted as 2:b. The second tuple was correctly predicted as 1:a. The third tuple was correctly predicted as 1:a.Tthe fourth tuple was correctly predicted as 1:a. The fifth tuple was incorrectly predicted As 1:a. but its actual class is 2:b. In 1:a, "1" refers to first class and "a" refers to diabetes patient.

In 2:b, "2" refers to second class and "b" refers to normal patient. The instances, actual, predicted, error and prediction are parameters of predictions on test split. Here, The instances refers to tuple of dataset. The actual is refers to the class label of tuple. The predicted refers to the class label predicted by the classifier.Error will be occurred when the classifier predicts incorrect class labels. Prediction is the value of the class label that were predicted by the classifier.

Table 3. Performance of the Classifiers

Parameters	**Naïve bayes**	**Multilayer perceptron**	**IBK**
Correctly classified instances (%) (in Test set)	100	88	88
Incorrectly classified instances (%) (in test set)	0	1	1
Kappa statistic	1	0.7692	0.7692
Mean absolute error	0.0234	0.1334	0.1347
Root mean squared error	0.0689	0.3221	0.3245
Relative absolute error (%)	4.7989	27.3272	27.5862
Root relative squared error	13.7901	64.4819	64.9681
Coverage of cases (0.95 level)(%)	100	88.8889	88.8889
Mean	55.5556	61.1111	50

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

rel.region Size (0.95 level) (%)			
Total number of instances	9	9	9
Time taken to build model(Second)	0.02	2.24	0

Correctly a classified instance is the set of tuples that are correctly classified by the classifier. Incorrectly classified instances are the set of tuples that are incorrectly classified by the classifier. Mean absolute error, root mean squared error, relative absolute error and root relative squared error-these are the measures of predictor error. Total number of instances is the set of tuples or records in a dataset.

a. Naive Bayes : Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	a
	1	0	1	1	1	1	b
Weighted Avg	1	0	1	1	1	1	

b. Multilayer Perceptron: Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.25	0.833	1	0.909	0.96	a
	0.75	0	1	0.75	0.857	0.917	b
Weighted Avg	0.889	0.139	0.907	0.889	0.886	0.941	

c. IBK: Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0.25	0.833	1	0.909	0.96	a
	0.75	0	1	0.75	0.857	0.857	b
Weighted Avg	0.889	0.139	0.907	0.889	0.886	0.833	

Table4. Average of the Classifiers

Algorithms	**Accuracy**
Naïve bayes	100%
Multilayer perceptron	88%
IBK	88%

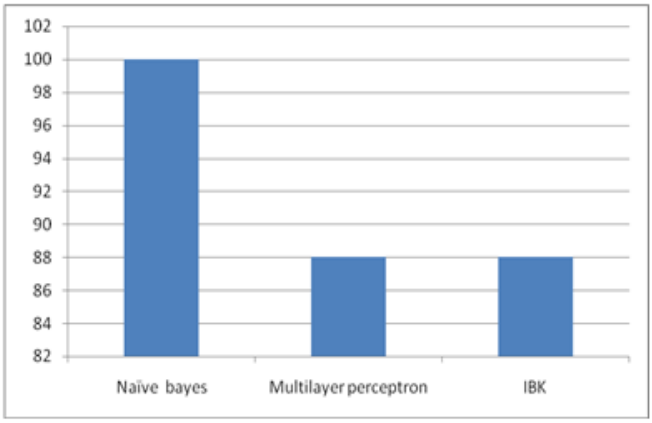

Fig.3. Prediction Accuracy

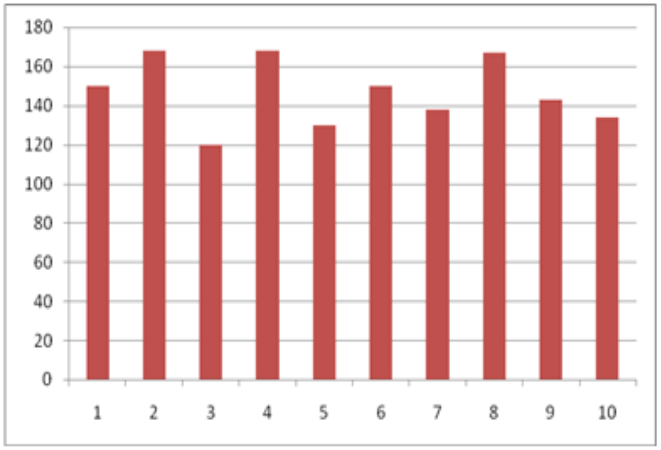

Fig. 4 class label prediction

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

V. CONCLUSION

The classifiers has been correctly predict the class labels a and b. The accuracy of the classifier will be efficient, when the classifier will correctly predict the instances of the dataset. After analysing the experimental results, conclude that each algorithm provides best result for our research.

The accuracy of the Naive bayes is 100%. The accuracy of the Multilayer perceptron is 88%. The accuracy of the IBK is 88%. So the Naive bayes algorithm gives best accuracy compared to IBK and Multilayer perceptron. In future, different data mining techniques can be used for predicting the different types of diseases.

REFERENCES

[1] Azwa Abdul Aziz, Nur Hafieza Ismail, Fadhilah Ahmad, "Mining Students' Academic Performance", Journal of Theoretical and Applied Information Technology, Vol. 53 No.3,31 st July 2013 , ISSN: 1992-8645

[2] Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás, "Data Mining Algorithms to Classify Students"

[3] A.A. Aziz, N. H. Ismail, & F. Ahmad, "Mining Students' Academic Performance", Journal of Theoretical and Applied Information Technology, vol. 53(2013), no. 3, 485–495.

[4] Han, J. and Kamber, M., "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Publishers, 2006.

[5] Jigna Ashish Patel , "Classification Algorithms and Comparison in Data Mining" , International Journal of Innovations & Advancement in Computer Science, IJIACS, Volume 4, May 2015, ISSN 2347 –8616