ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Comparative Study on Classifiers Using Diabetes Data Set.

B Santhi, and K Gokulnath*.

Department of Information & Communication Technology, SASTRA University, Tanjore, INDIA

ABSTRACT

Diabetes Mellitus is one of the most widespread diseases in all age group and commonly referred as Diabetes. It contributes several complications such as nerve and blood vessel damage, heart problems, and a higher risk of kidney malfunctioning. Data Mining, being the foremost analyzing technique used by researchers provides effective results in an early diagnosis of diabetes. The Data Set used for the diabetes data analysis is Pima Indians Dataset with 768 Samples. This paper compares different classifiers and identifies the best of them. This experimental study reveals Naïve Bayes outperforms than J48.

Keywords: Data Mining, Diabetes, Classification, J48, Naïve Bayes, WEKA

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Data Mining is the operation of extricating required data from abundant information. This has resulted in exploring hidden pattern from huge repository. Data Mining is a major evolution in the development of mining tools for analyzing data. It is an integrative field of combining Database Management Technology, Artificial Intelligence, Statistics, Soft-Computing and Machine Learning techniques. It provides higher accuracy in diagnosis and efficient patient treatment in the field of medicine. There are numerous Data Mining techniques available for analyzing the data. They have some phases for understanding the business requirements and data, exploring the data, modeling, examining, and deployment [1].

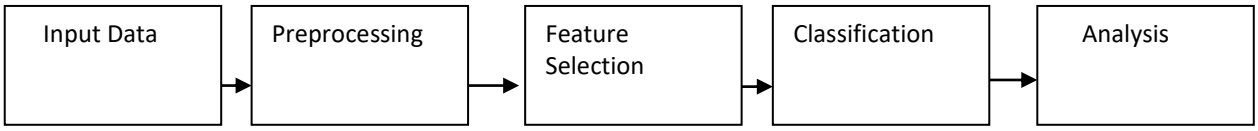


Fig 1: Workflow of Data Mining System (DMS)

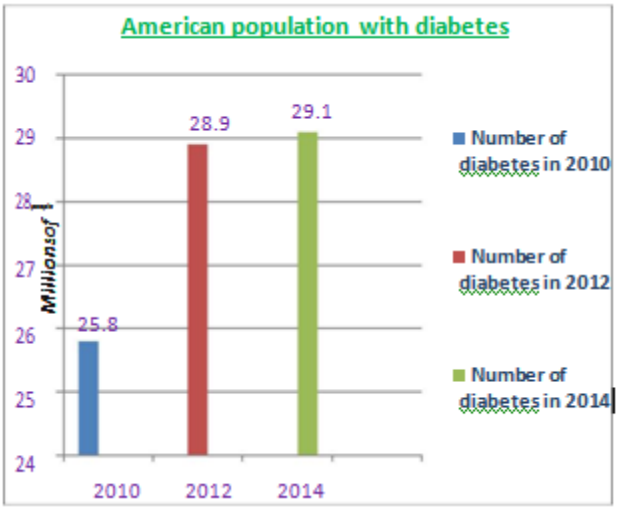
One among the most prevailing diseases in all racial and ethnic groups in the world would be Diabetes Mellitus, which can be fatal when left untreated. It is a metabolic disorder begot by deficient or nil creation of the hormone insulin by the pancreas. The insulin is the hormone required for directing the glucose into body cells. Some Changes occur in the creation of insulin, which will directly increase the blood sugar level. This production of higher insulin can impair the working condition of tissues, cells, nerves, blood vessels and organs of our body. It can be classified as three types of diabetes which are specified below and they would be Type-1, Type-2 and Gestational Diabetes [2].

The main goal is to use Data Mining techniques to explore hidden patterns that find the perfect classification of diagnosis of diabetes.

RELATED WORKS:

The recent survey in the CDCP (Center for Disease Control & Prevention) US (2014) shows that US population with diabetes is approximately equal to 9.3% which holds 29.1 million in overall population (i.e) every 1 people out of 11 people have diabetes and 1 out of 4 does not realize that they possess with diabetes[12]. Also, the survey includes results about prediabetes. Totally 86 million of the population possess prediabetes which is higher than 1 among 3 adults and 9 from 10 humankind do not realize that they do have prediabetes. In the overall population, 15-30% are having prediabetes.

Fig 2: Survey of Diabetes counts in the United States



According to the study, the body that does not create sufficient insulin at whichever age and cannot be prevented is called as “Type-1-diabetes”. It is sometimes said to be “Juvenile-Onset-Diabetes”. The human body which could not utilize insulin regularly can be forestalled in most cases. It can develop at any age is called as “Type-2-Diabetes” [3]. It is sometimes said to be “Adult-Onset-Diabetes". In adults, approximately only 5% of identified cases of diabetes result as Type-1. The diagnosed person to have prediabetes, when their glucose measure in blood is greater than the usual measure but not as peak amount to be recognized as type-2 diabetes [4]. People aging 20 and above in the United States, 2012 results the detected and undetected level of diabetes.

According to the National Health and Nutrition Examination of Survey in 2012 US data, they are classified into age and gender. By classifying with age, in 20 years and above there are 28.9 millions of diabetes and 12.3 percentages with diabetes in the overall population. Briefly, under 20 to 44 age, 4.3 millions of diabetes and 4.1 percentages with diabetes, fewer than 45 to 64 age, 13.4 millions of diabetes with 16.2 percentages with diabetes and above 65, 11.2 millions of diabetes and 25.9 percentages with diabetes in the overall population. By gender classification, Men have 15.5 millions of diabetes and which is 13.6 percentages with diabetes in total population and women have 13.4 millions of diabetes and which is 11.2 percentages with diabetes in the total population.

Classification of a model for diabetes diagnosis is the bedrock in many ongoing types of research for the past decades. The researchers were attentive for a long time by using Statistical data mining tools and thus by improving analysis of required data from large datasets. Most of them follow Clustering Algorithms and Neural Networks.

Diagnosing, Predicting and classifying the diseases are some of the applications which provide successful results using data mining tools. The abundant medical data requires the need of extracting useful knowledge using strong data analysis tool like WEKA. The content of biological data grows rapidly and analyzing them needs such powerful tool to explore, extract, organize, interpret and utilize the information. The impact of bioinformatics and data mining leads a significant role in diagnosing many diseases.

DATASET:

The main purpose is to predict whether the patient has been affected or not, using the data mining tools and with the available medical dataset. The proposed analysis uses the “Pima Indian Diabetes Data Sets” [14] acquired from the MLDatabase repository of UC, Irvine. The dataset was sampled from a huge set of data available in the NIDDK. In this dataset, it includes patients of Pima-Indian women inhabiting near Phoenix, AZ (USA) with at least of 21 years old. By analyzing the correlation between Plasma Glucose and Class attributes, we can easily predict the development of diabetes nearby future. A Higher value of Plasma-Glucose creates more chance to develop diabetes and with least chance of developing diabetes in future by having low Plasma-Glucose. This Data Set for the diabetes data analysis comprises with 768 Samples. After the test, patient response takes the value ‘0’ or ‘1’, in which ‘0’ determines the negative value and ‘1’ determines positive value [5].

All the attributes are numeric-valued and they are described as No. of times pregnant, Sugar level, Diastolic BP, Thickness in skinfold (mm), Serum with insulin for 2 hours, Body mass index(kg/ , Pedigree fn. for diabetes, Duration (yrs) and Class (either ‘0’ or ‘1’) [6].

By implementing data processing technique, J48 Algorithm is used on the dataset so that the data are classified into “Tested-Positive” and “Tested-Negative” based on its result using WEKA data mining tool.

ISSN: 0975-8585

METHODS AND MATERIALS

Preprocessing:

In Data Selection method, it is implemented to find the errors such as missing values, wrong content, and inconsistency of data. In the data analysis stage, it computes the data to get the required results by analyzing the datasets using a special tool like WEKA. Fig.3, Shows Preprocessing technique includes transformation which performs

- **Restoring the missing values**
- **Normalizing the data**
- **Finding the erroneous values**

As data in the real world is dirty, incomplete and noisy, we need to perform data preprocessing technique [10]. In this method, it involves finding the errors and missing the value of data from the abundant dataset. Using Preprocessing, it becomes easy to return the missing values and correct the inaccurate data.

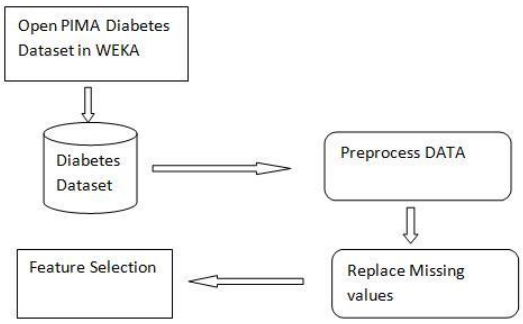


Fig 3: Data Preprocessing

Classification:

Classification is the method to find a set of models that elucidate by assigning an object to a certain class based on its similarity to previous examples of other objects. The classifier is developed to forecast the categorical labels. These labels classify a data item into any one of the inbuilt classes. Firstly, classification may indicate similarity to objects that are definitely members of a given class. All defined objects must undergo pre-classification (i.e.) the label should be understood. It would be assumed that every sample belongs to a predefined class. Secondly, classification deals with the model usage. It is only for classifying succeeding concealed objects [7].

Pima set is loaded in WEKA. Using preprocessing technique, missing values are filled. Then the preprocessed data is classified with J48 & Naïve Bayes. From the output, it is easy to conclude the better accurate classifier algorithm.

Classifier:

Naïve Bayes

It is easy to implement and good results are obtained in most of the cases. But practically, in most of the cases, it has low accuracy and dependencies exist among variables. So, Bayesian Belief Networks is used to deal with the existence of dependencies [8].

There exist the percentages splits of 66% and 80% are applied in Naïve Bayes. The Naïve Bayes algorithm gives the following correctness results for the given dataset. Through Naïve Bayes Classification Algorithm, we obtain few instances as correctly classified, yielding 84.29% accuracy.

In table 1, M1 is described as Precision, M2 is described as Recall, and M3 is described as F-Measure [9].

ISSN: 0975-8585

J48

J48 classifier algorithm is used to classify whether a patient has tested positive or negative diabetes with a decision tree structure. It is the execution of the ID3 algorithm (Iterative Dichotomiser 3) originated by the WEKA Project team. It was developed by Ross Quinlan. The decision tree produced by J48 can be utilized for classification.

The Same split percentage is allowed and J48 is checked. The J48 algorithm gives the following correctness results for the given dataset. Through J48 Classification Algorithm, we obtain few instances as correctly classified, yielding 81.61% accuracy

In table 1, M1 is described as Precision, M2 is described as Recall, and M3 is described as F-Measure

EXPERIMENTAL SETUP

Comparing both the results:

Confusion Matrix –

Predicted / Actual	**Naïve Bayes**		**J48**	
	A→ Observed	**B → Expected**	**A → Observed**	**B → Expected**
A → Observed	158 (i)	20 (ii)	159 (i)	19 (ii)
B→ Expected	21(iii)	62(iv)	29 (iii)	54 (iv)

Test Statistics	**Naïve Bayes**	**J48**
Kappa statistic	0.64	0.56
MAE	0.24	0.28
RMSE	0.34	0.36
RAE	53.01%	61.54%
RRSE	73.16%	77.15%
Total Number of Instances	261	261

In the table, the values represent following:

(i): Number of true forecasts in which the instances are Observed values. (ii): Number of false forecasts in which the instances are Expected values. (iii): Number of false forecasts in which the instances are Observed values. (iv) : Number of true forecasts in which the instances are Expected values.

ISSN: 0975-8585

Classifiers	TPR-True Positive Rate	FPR-False Positive Rate	M1	M2	M3	ROC Area Value
Naïve Bayes	0.88	0.25	0.883	0.88	0.88	0.89
	0.74	0.11	0.756	0.74	0.75	0.89
J48	0.89	0.34	0.84	0.89	0.86	0.86
	0.65	0.10	0.74	0.65	0.69	0.86

Table 1: Test Statistics

Terminologies of Test Statistics:

1. **Kappa Statistic:** It is a standard that collates observed Accuracy with Expected Accuracy (random chance).

2. **MAE-Mean Absolute Error:** an average of the absolute error between observed and forecasted value.

3. **RMSE-Root Mean Squared Error:**

It determines the differences between Sample value and population values. It could be easily estimated by a prototype or a predictor and also by the observed values.

4. **RAE- Relative Absolute Error:**
   It is the ratio of absolute error in measurement to the accepted measurement.

They can be derived and calculated using following formulae,

$$M1 = \frac{TP}{TP + FP}$$

$$M2 = \frac{TP}{TP + FN}$$

$$M3 = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

ISSN: 0975-8585

RESULTS AND ANALYSIS

J48 Classification Algorithm, we obtain some instances as correctly classified, yielding 81.61% accuracy. Naïve Bayes Classification Algorithm, we obtain some instances as correctly classified, yielding 84.29% accuracy.

Classifier Algorithms	Split for 66%	Split for 80%
J48	81.61%	80.52%
Naïve Bayes	84.29%	85.71%

When comparing the results of Naïve Bayes and J48 Classify Algorithm for Pima-Indian Diabetes dataset, both methods have a relatively least difference in accuracy rate. Though we apply the percentage split of 66 and 80 in Naïve Bayes technique, it produces a result of less error rate when related with the J48 algorithm. Though both the model is effective in the analysis of diabetes with the percentage split 66 and 80 for the Pima-Indian Diabetes dataset, Naïve Bayes shows greater accuracy rate. While using PIMA-Indian Diabetes dataset, Naïve Bayes performance is comparatively higher than J48.

CONCLUSION

The Pima - Indians Diabetes dataset has been used for the empirical purpose and it holds the data of patients with and without diabetes. Early prediction and diagnosis of diabetes are essential for treatment. By using data mining techniques the possibility of getting affected by diabetes is predicted early. Naïve Bayes algorithm mainly focused with probability and the J48 algorithm is focused on the decision tree. With the result, it is easy to conclude that Naïve Bayes have a higher number of the correctly classified instance and its accuracy for predicting correctly is also higher. This paper presents how Naïve Bayes and J48 classifier algorithm are used for detection of diabetes for curing. In future, the automatic diagnosis of diabetes can be developed.

REFERENCES

[1] Shivakumar.B.L and Alby.S, “A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes”, International Conference on Intelligent Computing Applications, IEEE 2014, pp 167-173.
[2] Sajida Perveena , Muhammad Shahbaza , Aziz Guergachib , Karim Keshavjeec, “Performance analysis of data mining classification techniques to predict diabetes”, March 2016, 82: 115-121.
[3] Arianna Dagliati, Lucia Sacchi, Carlo Cerra, Paola Leporati, Pasquale De Cata, Luca Chiovato, John.H. Holmes and Riccardo Bellazzi ,“Temporal data mining and process mining techniques to identify Cardiovascular risk associated clinical pathways in Type 2 diabetes patients”, IEEE 2014, pp 240-243.
[4] Sadri Sadi , Amanj Maleki , Ramin Hashemi , Zahra Panbechi and Kamal Chalabi, Comparison of Data Mining Algorithms in the Diagnosis of Type II Diabetes", International Journal of Computational Science & Applications (IJCSA), October 2015; **5**, 1-12.
[5] Velu.C.M and Kashwan.K.R, “Visual data mining techniques for classification of diabetic patients”, International Advance Computing Conference (IACC), 2013 IEEE, pp 1070-1075.
[6] Sonu Kumari, “A data mining approach for the diagnosis of diabetes mellitus”, Intelligent Systems and Control (ISCO), IEEE 2013, pp 373-375.

[7] Aiswarya Iyer, Jeyalatha.S and Ronak Sumbaly, "DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES", International Journal of Data Mining & Knowledge Management Process (IJDKP), January 2015;Vol.5,No.1, pp 1 – 14.
[8] Ranjit Abraham et.al, "A comparative analysis of discretization methods for Medical Datamining with Naïve Bayesian classifier", International Conference on Information Technology, IEEE 2016.
[9] Shaleena.K.P et.al, "Data mining techniques for predicting student performance", International Conference on Engineering and Technology, IEEE 2015.
[10] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition.
[11] http://www.diabetes.org/
[12] http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report- web.pdf