IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

Performance Analysis of Diabetes Disease using Classification Algorithms by WEKA

G. Paul Davidson[1*], D. Ravindran[2]

[1]Department of Computer Applications, Bishop Heber College, Tiruchirappalli, India
[2]Department of Computer Science, St. Joseph's College, Tiruchirappalli, India

Available online at: www.ijcseonline.org

Abstract— In Medical industry there are many diseases that makes a patient critical among them diabetes is one of the major disease that affect most of the people in early stage. Diabetes (or Diabetes Mellitus) is a group of metabolic diseases, chronic, in which there are high blood sugar levels and affects the body's ability to use the energy found in food over a prolonged period. Researchers are finding effective methods for the prediction of diabetes. The main goal is to analysis the performance of various data mining techniques in the diabetes dataset for efficient extraction of valuable patterns. For doing so WEKA software was used as a mining tool for diagnosing the useful pattern. The Pima Indian diabetes dataset are used for the analysis. The dataset was applied in various classification algorithms to analysis the performance to identify an effective model that predict diabetes disease. In this, the analysis is done by applying attribute evaluator to enhance the accuracy then applying Naive Bayes, Bayes Net, J48 and Random Forest and the performance are compared. Through this study, Naive Bayes Algorithm provides better classification accuracy, when compared with classification algorithms like Bayes Net, J48 and Random Forest.

Keywords: Diabetes, Health care, Naive Bayes, Bayes Net, J48 and Random Forest, WEKA.

I. INTRODUCTION

Healthcare industry contains very large and sensitive data. This data needs to be treated very carefully to get benefitted from it. There is need to develop some more accurate and efficient predictive models that helps in diagnosing a disease although it was revealed that diabetes mellitus is the diseases which becomes one of the global hazard. Diabetes mellitus referred to as diabetes, is a group of levels over a prolonged period.

This diseases that affect how the body uses blood sugar (glucose). Glucose is vital to the health because it's an important source of energy for the cells that make up your muscles and tissues. Diabetes mellitus is characterized by recurrent or persistent high blood sugar, and is diagnosed by demonstrating plasma glucose level. The main focus is to find an algorithm to diagnose diabetes. For this data mining techniques are used. Data mining, "a major way of creating knowledge", is a useful way of studying medicine, genetics, bioinformatics, education.

Data mining techniques can be classified into both unsupervised and supervised learning techniques. Unsupervised learning technique is not guided by variable and does not create a hypothesis before analysis. Based on the results, a model will be built. A common unsupervised technique is clustering. Supervised learning technique requires the building of a model that is used in prior performing analysis. Usually Supervised learning techniques that are used in both medical and clinical research are Classification. By using the WEKA tool the diabetes datasets are passed through various classification algorithm such as Naive Bayes, Bayes Net, J48 and Random Forest and the result are compared to analysis the best performance of those algorithms.

II. LITERATURE STUDY

Usma el al[16], mention the accuracy can be increase by improving the performance of the data, the algorithms or even by algorithm tuning. To enhance the accuracy improve the pre-processing phase. Applying bootstrapping resampling technique on the PIMA dataset will increases the accuracy of almost all classifiers but the decision trees leads over others. It is also concluded that the accuracy of a model is highly dependent on the dataset. *Yasodha et al*[11], discuss the uses of classification on diverse types of datasets to identify whether a person has diabetic or not. The diabetic patient's data set has 249 instance and 7 attributes gathered from hospital warehouse. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study by using WEKA tool the datasets are classify and assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random

Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others.

Saravananathan et al[8], mention with a proper data preprocessing technique can get better the accuracy of the classifier. The function of data normalization had noticeable impact on categorization performance and considerably enhanced the performance of J48. The performance of kNN algorithm has minimum accuracy. Based on the parameters taken for analysis, the performances of the four algorithms are analyzed. The results show that the performance of J48 technique is superior to the other three techniques for the classification of diabetes data.

Aiswarya et al[2], aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split.

VelidePhani et al[17], discuss that the various data mining techniques are applied to the datasets in order to get the accuracy and time that the algorithm found the result and also concluded that J48(C4.5) with 9 attributes had shown accurate and better performance with least time taken for analysis of Diabetes data.

III. FRAMEWORK

In the proposed model, we have used a classification model with pre-processed dataset by applying Attribute selection method and InfoGain algorithm for the Diabetics Dataset. Some classifier like Naïve Bayes, Bayes Net, J48 and Random Forest are applied in those diabetes Datasets. The main aim is to analysis the performance of those algorithms by analyzing the accuracy of grouped class. The framework is shown in Figure 1.

Phases

The framework is composed of the following important phases:

- Dataset Selection (PIMA Indian Diabetes Dataset)
- Data Preprocessing
- Feature extraction through InfoGain Attribute Eval using ranker Method
- Applying Attribute Selection filter
- Learning by Classifier (Training) i.e. NaïveBayes, Bayes Net, J48, Random Forest
- Achieving trained model with highest accuracy
- Analysis the model and predict.

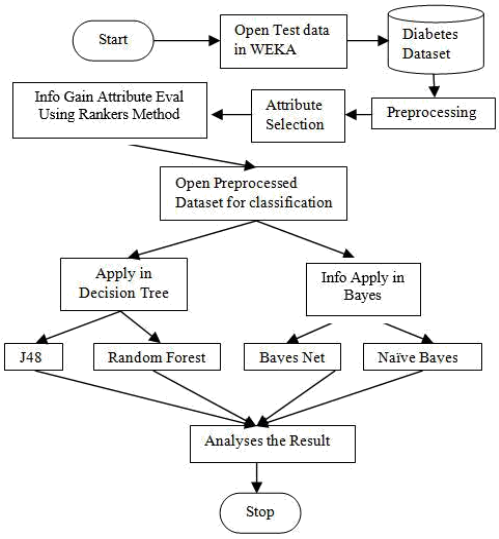

Fig1. The Proposed Framework

Dataset Selection (Diabetes Dataset)

The selection of data is a process in which the most relevant data is selected from a specific domain to derive values that are informative. In this Process, diabetes dataset are used that have eight attributes used to predict the symptom of diabetes in a female patient. This dataset was obtained from UCI repository. On the basis of historical information stored in the dataset such as age, body mass index, blood pressure and number of times pregnant the classifiers are trained for making decision whether diabetes test for an individual is positive or negative. The PIMA diabetes dataset only represents the Indian national females who are at least 21 years old. All of the attributes are of numeric-valued continuous data type. PIMA Indian Diabetes Dataset from UCI repository contains 768 instances. The PIMA dataset is converted from CSV to ".ARFF" format accepted by WEKA 3.8.2.

WEKA Tool

WEKA 3.8.2 is used in this study. WEKA stands for the Waikato Environment for Knowledge Analysis and this tool is developed and distributed freely by the University of Waikato, New Zealand. WEKA is one of the most famous tool for data processing and data analysis. Since WEKA software has been written in Java language, therefore, it runs on almost every platform. It consists of variety of machine learning algorithms and is capable to solve a multitude of data mining and machine leaning problems. WEKA supports many data mining tasks such that regression, classification, prediction, feature selection and visualization. WEKA allows us to create, run, modify and analyze experiments. The advantages of WEKA include its free availability, portability, a broad collection of data preprocessing and modeling techniques and the friendly graphical user interface makes it easy to use.

Data preprocessing

Data preprocessing is a technique of machine learning that comprises of converting raw data into a logical or comprehensible format. Data preprocessing is a conventional technique of eliminating problems which are known as noise. Preprocessing involves certain activities like data cleaning, integrating the data, transformation of data, data reduction, data discretization and data cleaning. Here the dataset is checked for duplicate values, missing values and type miss-matches etc.

All these inconsistencies are eliminated from this dataset, in the phase called data preprocessing phase. It is important to clean the dataset before training it on a classifier in order to better learn the hidden patterns in the dataset.

Information Gain Selector:

Information gain ratio is to gain the basic information. It is used to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute. Information Gain is also known as Mutual Information.

Attribute Selection Filter:

Attribute selection is so important that WEKA dedicates a separate package to host related files. To perform attribute selection, three elements are required. One is search method, and the second is evaluation method. Both elements need to be initiated and defined in a container class Attribute Selection. The third element is data.

Classifiers

A classifier is a tool in machine learning that proceeds a group of data demonstrating the objects we need to classify and tries to forecast which class the new data belongs to. The classification objective set for this study is to achieve enhanced accuracy by using Naïve Bayes, Decision Trees to determine which one suits the most for diabetes classification technique. The techniques used are Naïve Bayes, Bayes Net, J48 and Random Forest.

Bayes Net:

A Bayesian network[3] is a representation of a joint probability distribution of a set of random variables with a possible mutual causal relationship. The network consists of nodes representing the random variables, edges between pairs of nodes representing the causal relationship of these nodes, and a conditional probability distribution in each of the nodes. The main objective of the method is to model the posterior conditional probability distribution of outcome variables after observing new evidence. Bayesian networks may be constructed either manually with knowledge of the underlying domain, or automatically from a large dataset by appropriate software.

Naive Bayes:

It is a simple technique for constructing classifier models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle. All naive Bayes[10] classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

J48 Decision Trees:

The J48[7] Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

Random Forest:

Random forest[12], also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. It is to generate multiple small decision trees from random subsets of the data. Each of the decision tree gives a biased classifier. They each capture different trends in the data.

Ensemble methods use multiple learning models to gain better predictive results. In the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

IV. EXPERIMENT AND RESULT

For experimentation PIMA Indian diabetes dataset is used. The PIMA dataset, have two class in individual patient having tests either positive or not. The dataset consists of 768 total instances and nine attributes. After preprocessing the data instances are reduced. Attribute Selection Filter is used to reduce the dimensionality of dataset. By applying Info Gain Attribute Evaluation on all the attributes, it returned five attributes to be used for training the classifiers. Then

applying filter with no replacement that disables the data to be replicated. The classifiers are applied. The naïve Bayes, Bayes Net,J48, Random Forest classifiers are applied one by one on the same data. The classification results are evaluated by comparing them in terms of correctly classified and incorrectly classified instances. Here only three performance measure that includes accuracy, precision and recall are taken into account.

The formula used to calculate the Accuracy is mentioned below:

$$\textbf{Accuracy} = \frac{TP}{TP+FP} \quad \ldots\ldots (1)$$

Precision indicates the number of True Positives divided by the number of True Positives and False Positives. Hence, it shows the number of positive predictions divided by the total number of positive class values predicted. Precision is also termed as the Positive Predictive Value (PPV). The formula is mentioned below:

$$= \frac{TP+TN}{TRUE+FALSE}$$

$$\textbf{Precision} \quad \ldots\ldots (2)$$

Whereas recall indicates the number of True Positives divided by the number of True Positives and the number of False Negatives. Hence it is the number of positive predictions divided by the number of positive class values in the test data. Recall also sometimes titled as Sensitivity or the True Positive Rate. The formula is mentioned below:

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

$$\textbf{Recall} \quad \ldots\ldots (3)$$

Performances of each classifier are measured in these terms by using equation 1, 2 and 3.
The final results are shown below:

Table: 1 Confusion Matrix for Bayes Net

Bayes Net		Predicted	
		Negative	Positive
Actual	Negative	426	74
	Positive	114	154

Accuracy=75.52
Precision=78.88
Recall=85.2

Table: 2 Confusion Matrix for Naïve Bayes

Naïve Bayes		Predicted	
		Negative	Positive
Actual	Negative	438	74
	Positive	111	157

Accuracy=77.47
Precision=79.78
Recall=87.6

Table 3: Confusion Matrix for J48

Classifier	TP	FN	FP	TN	Accuracy	Precision	Recall
Bayes Net	426	74	114	154	75.52	78.88	85.2
Naive Bayes	438	62	111	157	77.47	79.78	87.6
J48	426	74	119	149	74.86	78.16	85.2
Random Forest	415	85	109	159	74.73	79.19	83

Accuracy=74.86
Precision=78.16
Recall=85.2

Table: 4 Confusion Matrix for Random Forest

J48		Predicted	
		Negative	Positive
Actual	Negative	426	74
	Positive	119	149

Accuracy=74.730
Precision=79.19
Recall=83

Table: 5 Comparison of all Data in all Classifier

Random Forest		Predicted	
		Negative	Positive
Actual	Negative	415	85
	Positive	109	159

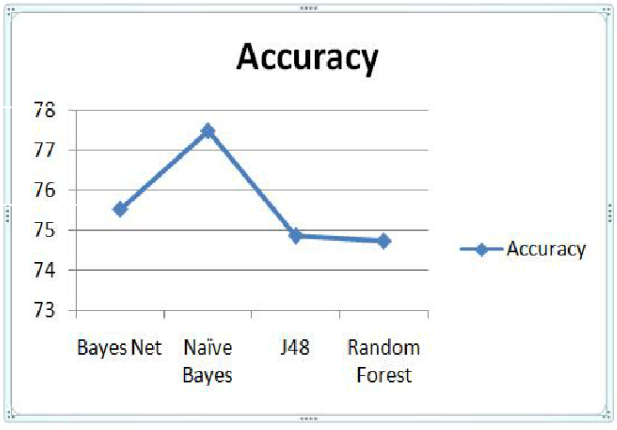

Fig2. Graph for Accuracy

Table: 6 Comparison with various authors

Author	Classifier/Method	Dataset	Accuracy(%)
Usma et al (2017)	Naïve Bayes	PIMA	74.89
	J48		94.44
Yasodha et al (2011)	Bayes Net	PIMA	66.2
Saravananathan et al(2016)	J48	PIMA	67.15
Aiswarya et al(2015)	J48	PIMA	74.8
Velide et al (2014)	Naïve Bayes	PIMA	55.85
	J48		68.58
Proposed work	Bayes Net	PIMA	75.52
	Naïve Bayes		77.47
	J48		74.86
	Random Forest		74.73

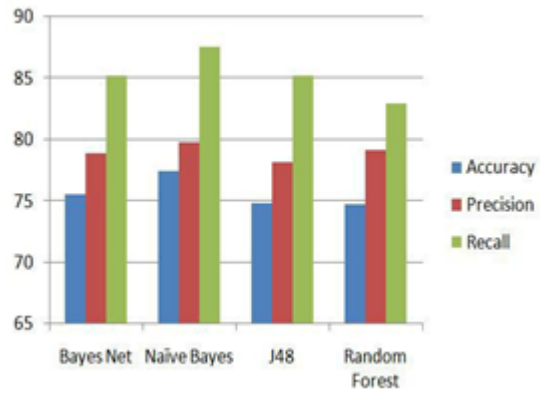

Fig3. Bar Chart for Performance Comparison

Comparison of the Performance:

The proposed experimental results are compared other researchers work. By comparing our proposed work with other author work all method shows better accuracy except. In the proposed work by using InfoGain Attribute evaluator in all classifier methods Naïve Bayes show high accuracy when compared to other three algorithms. So the same Naïve Bayes when compared with other authors our result is much better than others. So Naïve Bayes Shows best accuracy.

V. CONCLUSION

Diagnosis of Diabetes is Important in the Medical World. Diagnosis of Diabetes in early stage is useful to avoid risk factor in the patients. In this work, it is shown that various classification algorithms like Bayes Net, Naïve Bayes, J48 and Random Forest. The performances of the result are compared. Since all the result shows little difference only. But among them Naïve Bayes has better Accuracy. So it is concluded that Naïve Bayes show Better accuracy by using Info Gain Attribute Evaluator and Attribute Selection Filter. In our Future Work the same algorithm can be used with various dataset or the same algorithms can be compared with other classification algorithms using any other filter method.

REFERENCES

[1] Asma A Aljarullah. Decision tree discovery for the diagnosis type-2 diabetes. International conference on innovation in information technology. 2011; p. 303-7.
[2] Aiswarya Iyer, Jeyalatha S and Sumbaly Ronak. Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process. 2015; 5:1-14. 2.
[3] "Bayes Net", Wikipedia, Aug 2018.
[4] ChaitraliDangare, S. and SulabaApte,S.Improved study of disease prediction using data mining classification techmiques. Int.J.Comp.Appl., 2012,47(10):75-88.
[5] Global Diabetes Community, http://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html
[6] Ianchao Han J, Juan C Rodriguze, Beheshti Mohsen. Diabetes Data Analysis and Prediction model discovery. Second International conference on future generation com- munication and networking. 2011; p. 96-9. 13.
[7] "J48", Wikipedia, March 2018.
[8] K. Saravananathan and T. Velmurugan "Analyzing Diabetic Data using Classification Algorithms in Data Mining" Indian Journal ofScience and Technology, Vol 9 (43) | November 2016 | www.indjst.org
[9] Maniya Hardik, Mosin I Hasan, Komal P Patel. Comparative study of Naive Bayes Classifier and kNN for Tuberculosis. International Journal of Computer Applications. 2011; p. 22-6.
[10]"Naïve Bayes", Wikipedia, March 2018.
[11]P.Yasodha and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WekaTool", International Journal of Scientific & Engineering Research, vol. 2, no. 5, 2011.
[12]"Random Forest", Wikipedia, March 2018.
[13] Sankaranarayanan.S and Dr Pramananda Perumal.T, "Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies", World Congress on Computing and Communication Technologies, 2014, pp. 231-233
[14]Sonu Kumari and Archana Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus", Proceedings of71hInternational Conference on Intelligent Systems and Control (ISCO 2013)
[15]Stutz J., P. Cheeseman. (1996) Bayesian classification (autoclass): Theory and results. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press
[16]Uswa Ali Zia, Dr. Naeem Khan "Predicting Diabetes in Medical Datasets Using Machine Learning Techniques" International Journal of Scientific & Engineering Research Volume 8, Issue 5, May-2017, ISSN 2229-5518
[17]Velide Phani Kumar and Velide Lakshmi. A Data Mining Approach for Prediction and Treatment of diabetes Disease. International Journal of Science Inventions Today. 2014; 3:73-9.

