PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR DIABETIC PREDICTION USING PIMA- INDIAN DATASET

Dr. L. Arockiam[1], A. Dalvin Vinoth Kumar[2], S. Sathyapriya[3],
Associate Professor[1], Assistant Professor[2], M.Phil. Scholar[3]
Department of Computer Science[1,3], School Of CSA[2]
St. Joseph's College (Autonomous), Tiruchirapalli, India[1,3] , REVA University, Bangalore, India[2] .

Abstract : Data Mining is a process of collecting, extracting the data from various data warehouse and summarizing the data as a useful one. Essentially, data mining is referred to as "Knowledge Discovery from Data" (KDD) that is an extraction of knowledge automatically or in a convenient way . The predictive analytics is one such branch of advanced analytics in data mining, which is used to predict the future events. The predictive analytics uses different techniques such as machine learning, statistics, data mining and AI for analyzing massive data. This paper provides a review on predictive analytics and elaborates the predictive techniques with their application. Using the Pima Indian diabetes dataset, the prediction for diabetes is done with the help of various classification algorithms such as J48, Naïve Bayes and KNN. The accuracy of the algorithms is compared and found that J48 algorithm gives the highest accuracy.

IndexTerms **- Classification, Data mining, J48, KNN, Naïve Baiyes, WEKA.**

I. INTRODUCTION

Data mining is the process of discovering the knowledge from various data sets. which also defined as "Knowledge Discovery in data bases". Data mining techniques are applied on large volume of data for finding hidden models and relationship among the patterns which are helpful in decision making [1].Data mining techniques can be classified as descriptive model and predictive model [3]. The descriptive model represents the data in a brief form and applied to discover patterns in data and to analyze the relationship between attributes. The descriptive techniques include association mining, sequence discovery, clustering and summarization. The predictive model operates by predicting the values of unknown data from the known results. This includes regression, classification, analysis and prediction. The figure 1 depicts the Data mining techniques.

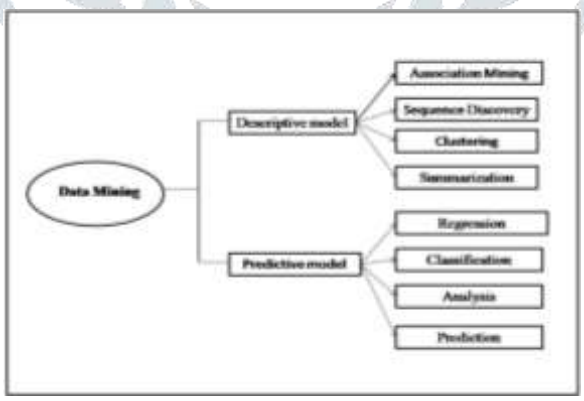

Figure 1: Various techniques of data mining

II. EXPLANATION OF DATA MINING TECHNIQUES

Classification

Classification categorizes data into one of the predefined classes. It has two processes. First one examines the objects and builds a model using training data which describes predetermined set of data classes. Secondly , the objects are assigned to a predefined class and classification techniques are applied. It mostly uses classification techniques such as Bayesian classifiers, Support Vector Machines, K-Nearest Neighbor, decision trees and neural networks.

Regression

Regression is the oldest and most popular statistical technique used for numeric prediction. This is used to map a data item to a real valued prediction variable. Regression analysis is used to identify the relationship between independent variables and dependent variables.

Time Series Analysis

Time series analysis encompasses methods and techniques for analyzing time series data in order to extract meaningful statistics. The values usually are obtained at uniform time intervals (hourly, daily, weekly, etc.).

Prediction

Prediction is used to predict future data which are relevant to past and current data. A few applications of prediction include speech recognition, machine learning, and pattern recognition.

Clustering

Grouping of objects is where similar objects exist in the same cluster and dissimilar objects exist in different clusters is called Clustering. It is also known as unsupervised classification. The similarity is calculated using Euclidian distance. The different types of clustering include, Hierarchical clustering, Partition clustering, Categorical clustering, Density based clustering and Grid based clustering.

Summarization

Summarization is also called as characterization or generalization. It summarizes a subset of data. The information about the database is collected by retrieving portions of the data. The resulting information is a set of aggregate information.

Association Rule Mining

Exploration of association rules between attributes in a transactional database is done and the association rules are used to find the frequency of items occurring together. These extracted rules are defined based on user defined minimum support value with minimum confidence value. This enables effective decision making. The algorithms used for association mining are Apriori algorithm, FPgrowth algorithm, Partition algorithm, Pincer-search algorithm and Dynamic Itemset Counting algorithm.

Sequence Discovery

Sequential discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. The patterns identified are most likely to have similar data and the relationship is based on time.

III. RELATED WORKS

Jan Andrzej Napieralski [7] discussed the different algorithms for predicting the items given in database. The process of prediction was accomplished by using various statistical methods. In addition to that, the appropriate preprocessing methods were implemented.Later the statistical method was applied.At the end of the experiment, probability of data prediction was calculated by using logistic regression and R programming language. Furthermore, additional mathematical methods were used while pre-processing the data on the dataset for better performance.

Satr et al [8] presented the review of decision tree data mining algorithms such as CART and C4.5. This paper provided the comparative study of both CART and C4.5 algorithm. Commonly decision tree algorithm can be used to predict the target value of its inputs. From the experiment, it is proved that the C4.5 algorithm is better than the CART algorithm.

Pragati et al [9] focused in the diagnosis of diabetes Mellitus using data mining techniques and analyzed k-fold cross validation, classification method, class wise K- Nearest Neighbor[CKNN], Support Vector Machine [SVM], LDA Support Vector Machine and Feed Forward Neural Network, Artificial Neural Network, Statistical Normalization and Back propagation methods for diabetic diagnosis. And presented that, SVM provided better accuracy on diabetic dataset.

Priyanka Chandrasekar et al [10] presented the method for improving the accuracy of decision tree mining with preprocessing data. Preprocessing method presented the benefits of classification accuracy performance tests. In this paper, the supervised filter discretization was applied with J48 algorithm. The process of proposed model classified the data by both

training and tested dataset. Classification accuracy was improved by entropy-based discretization method. Finally, the performance of this approach was compared with the J48 algorithm.

SwaroopaShastri et al [11] discussed about the type 2 diabetes disease and predicted using data mining algorithms whether diabetic patients had the diabetic kidney disease (DKD). The DKD patients information was collected who were affected by diabetic and prediction was done based on given attributes. The AES algorithm and Apriori algorithm were used for correlating and mining the set of items from the database. It established the correlation between diabetes and kidney disease patients. It helped the doctors to suggest the best medicine to the patient.

Various Data mining techniques, tools and data sets surveyed are presented in table 1.

Table 1: comparison between classification algorithms with various applications

Techniques & Algorithm	**Dataset & Tool**	**Parameter**	**Application domain**
J48 , LAD tree , MCC,MAE,RAE,RMSE,RRSEC	NBSS & LUB dataset (National Bureau of Soil Survey and Land use Planning. WEKA Tool	Accuracy, sensitivity, specificity	Agriculture [12]
J48 , SMO , Naïve Bayes , Multi Layer Perception	Complete B Blood count data set WEKA tool	Performance Accuracy Precision, Recall, True Positive rate, false Positive rate, F- measure	Medical [13]
K Nearest Neighbor, Decision Tree	Diabetic Dataset	Time Reduction, reduced cost, Accuracy	Medical [14]
J48, Naïve Bayesian	Diabetic dataset from medical college hospital. WEKA tool	Accuracy , Productivity	Medical [15]
J48, K-Means, Clustering, Decision Tree, Classification algorithms	Diabetes data set. WEKA TOOL	Accuracy	Medical [16]
J48, Deceision Tree, Multilayer perception, Naïve Bayes, Sequential Minimal Optimization	Turkey student evaluation records	Accuracy	Education [17]
PSO (Particle Swarm Optimization Algorithm), ANFIS (Adaptive Neuro Fuzzy Inference System), AGKNN (Adaptive Group Based K-Nearest Neighbor)	Diabetic Patients Record Dataset. MATLAB	Performance, Accuracy, Efficiency, Reduce Complexity	Medical [18]
Logistic Regression, Naïve Bayes	Heart patient dataset	Accuracy	Medical [19]
J48, Decision Tree algorithm , Meta – Technique	Private soil testing laboratory in pure(India)	Accuracy	Agriculture [20]
Artificial Neural Network,	Climate & Soil Dataset	Efficient	Agriculture

Back Propagation Training Method , Feed Forward Algorithm			[21]

IV. DATA DESCRIPTION AND ANALYSIS OF PIMA DATA SET IN WEKA

The data set collection is one of the important processes in data mining. The most relevant data is chosen from a particular domain for further analysis. The derived values can be more flexible and informative in that domain. In this study, PIMA Indian diabetic data set was used and it having nine attributes which are considered to predict diabetes. These sets of data obtained from UCI repository and the data set contains the basic knowledge about individuals such as age, BMI, BP and pregnant ladies, etc. The above mentioned attributes are numeric values with continuous data type. Totally, the data set having 768 instances, 9 attributes that have shown in table 2.

Table 2: Dataset Description

S.No	Attribute	Description	Maximum level
1.	Pregnancies	Total number of pregnant times	17
2.	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	199
3.	BloodPressure	Diastolic blood pressure (mm Hg	122
4.	SkinThickness	Triceps skin fold thickness (mm)	99
5.	Insulin	2-Hour serum insulin (mu U/ml)	846
6.	BMI	Body mass index (weight in kg/(height in m)^2)	67.1
7.	DiabetesPedigree Function	Diabetes pedigree function	2.42
8.	Age	Age (years)	81
9.	Outcome	Class variable (0 or 1)	-

In order to use WEKA tool, the data set available in .csv format was converted to .arff file format. Weka 3.6.13 is the latest version which is used in this study [26]. Weka consists of many machine learning algorithms which are capable to solve problems of data mining and machine learning. The converted data set from .csv is applied in weka for classification. The overview of weka environment after applied data set in it is shown in figure 2.

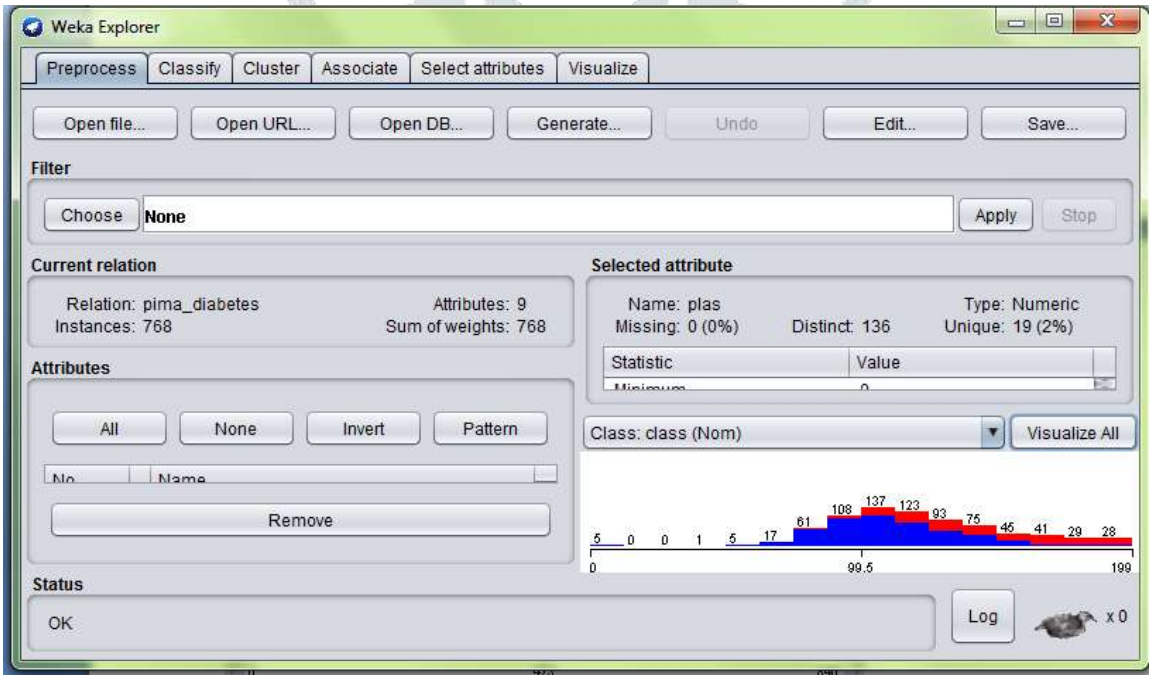

Figure 2: Overview of the weka tool environment

V. EXPERIMENTAL RESULTS

Pima Indian diabetes dataset is used for this study. The preprocessing techniques were applied on the instances of data sets. The Principle Component Analysis (PCA) is applied for reducing the dimensionality of dataset and it returned six attributes to be used for training the classifiers. The resample filter was used for omitting the replication of data. Hereby, the classifiers and cluster algorithms were applied to the Pima data set. The flow of attributes in Pima diabetic dataset is shown in figure 3.

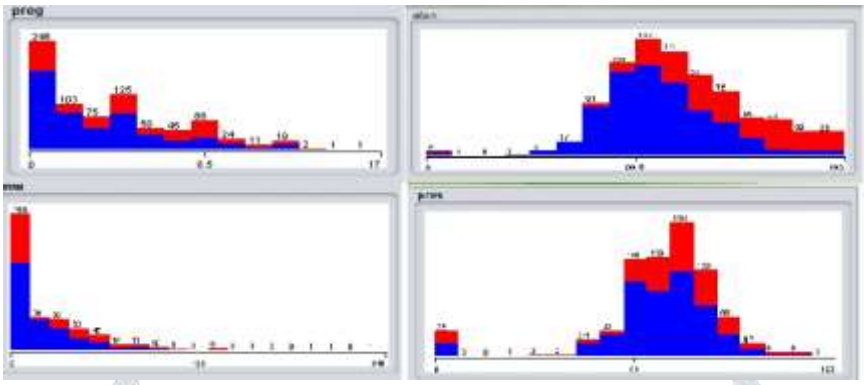

Figure 3: Process flow for attributes of PIMA Indian dataset

The results of classified instances are evaluated by comparing them in terms of Correctly Identified Correct Instances (CICI), Correctly Identified In-correct Instances (CIII), In-correctly Identified Correct Instance (IICI) and In-correctly Identified In-correct Instances (IIII). The most important operation in this work is to find accuracy, recall and precision. The f -measure and ROC were applied from this. F measure means average of precision and recall. In this study, three performances were considered viz Accuracy (Auc), Precision (Pcn) and Recall (Rcl). Auc is an arbitrary performance measure and deals with ratio of correctly predicted observation. If the class is balanced then the accuracy is best to measure. The formula used to calculate the accuracy is shown in equation 1,

$$\boldsymbol{Auc} = \frac{CICI+CIII}{TotalIdentifiedInstancese} \qquad (1)$$

Precision denotes the number of True Positives that are divided by the number of True Positives and False Positives. Therefore, it predicts the number of positive predictions divided by the total number of positive class values. Precision is also named as the Positive Predictive Value (PPV). The formula to calculate the precision is given in equation 2. Similarly, recall denotes the number of True Positives which are divided by the number of True Positives and the number of False Negatives. From this, the number of positive predictions divided by the number of positive test data class values. Recall is also known Sensitivity or the True Positive Rate. The formula to calculate the recall is given in equation 3. In this study, three algorithms such as J48, Naïve Bayes and KNN are used on the PIMA Indian dataset to classify the data. The comparison of precision and Recall in PIMA Indian Dataset for NB, J48 and KNN is shown in figure 4.

$$\boldsymbol{PPV} = \frac{CICI}{CICI+\text{IICI}} \qquad (2)$$

$$\boldsymbol{Rcl} = \frac{CICI}{CICI+\text{IIII}} \qquad (3)$$

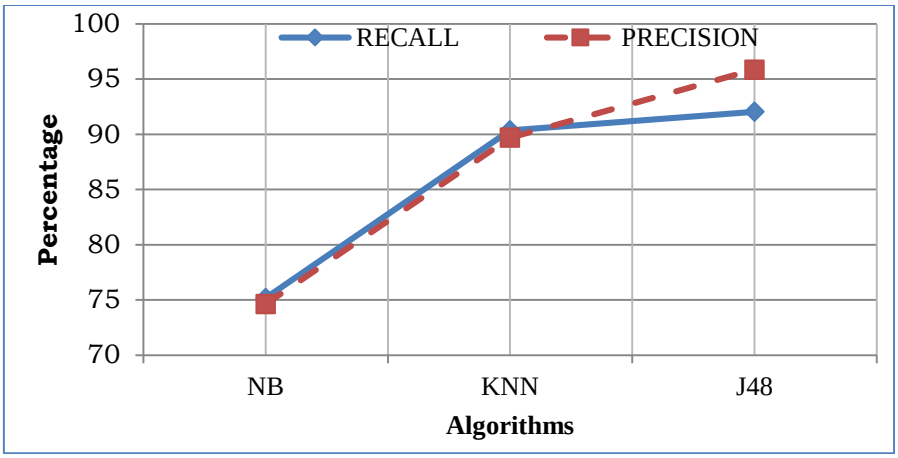

Figure 4: Precision and Recall for classification algorithm.

The accuracy is computed for three classification algorithms with PIMA Indian dataset and each accuracy is compared within themselves which is shown in the figure 5. It is observed that, J48 gives the higher accuracy (94.39 %) than Naïve bayes and KNN.

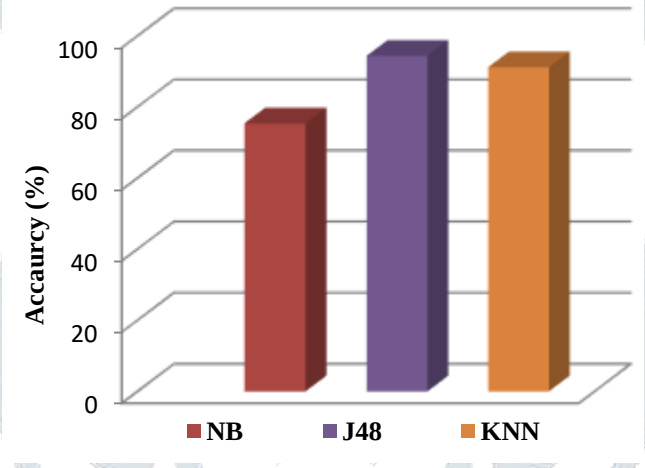

Figure 5: Accuracy for classification algorithms.

VI. CONCLUSION

Predictive analytic is the most essential and widely used technique. The process of predictive analytics is done with the help of various techniques such as data mining, machine learning and statistical tools. In this study, data mininig techniques was used for predicting diabetes disease which uses PIMA Indian diabetes data set. In order to diabetes disease was predicted by using three different algorithms like KNN, Naïve Bayes and J48. Finally, the result obtained from the experiment and J48 provided the better accuracy than other algorithms.

ACKNOWLEDGEMENT

This research work is financially supported by University Grants Commission, Government of India, under the Minor Research Project scheme. Ref. No.: F MRF-6517/16 (SER)/UGC).

REFERENCES

[1] J. Han and M. Kamber, ”Data Mining: Concepts and Techniques”, *Third Edition (The Morgan Kaufmann Series in Data Management System`s)*, 2000, 3rd Edition.

[2] K.S Deepikashri and Ashwinikamath, “Survey on Techniques of Data Mining and it’s Applications”, *International Journal of Emerging Research in Management &Technology*, Vol. 6, Issue: 2, 2017, pp: 198-201.

[3] Pradnya P. Sondwale, “Overview of Predictive and Descriptive Data Mining Techniques” *IJournals: International Journal of Software & Hardware Research in Engineering*, Vol.5, Issue. 4, 2015, pp: 53-60.

[4] Smita, Priti and Sharma, "Use of Data Mining in Various Field: A Survey Paper", *IOSR Journal of Computer Engineering*, Vol.16, Issue.3, 2014, pp: 18-21.

[5] Sayyed Muzammil Ali, Prof. Ms. R.R Tuteja," Data Mining Techniques", *International Journal of Computer Science and Mobile Computing*, Vol.3, Issue. 4, 2014, pp: 879 – 883.

[6] Brijesh Kumar Baradwaj, SaurabhPal" Mining Educational Data to Analyze Students Performance",(*IJACSA) International Journal of Advanced Computer Science and Applications*, Vol.2, Issue.6, 2011, pp: 63-69.

[7] Jan AndrzejNapieralski, " Statistical Methods for Data Prediction", *Department of Microelectronics and Computer Science, Lodz University of Technology*, 2016, pp: 20-24.

[8] SatbirKaur and Harjit Kaur," Review of Decision Tree Data mining Algorithms: CART and C4.5", *International Journal of Advanced Research in Computer Science*, Vol.8, Issue.4, 2017, pp: 436-439.

[9] Agrawal, Pragati, and Amit kumar Dewangan. "A brief survey on the techniques used for the diagnosis of diabetes-mellitus." *Int. Res. J. of Eng. and Tech. IRJET* , Vol. 2, Issue. 3, 2015, pp. 1039-1043.

[10] Priyanka Chandrasekar, Kai Qian, HossainShahriar and Prabir Bhattacharya," Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing", *IEEE 41st Annual Computer Software and Applications Conference*, 2017, pp: 481-484.

[11] Swaroopa Shastri, Surekha, Sarita, " Data Mining Techniques to Predict Diabetes Influenced Kidney Disease", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*,vol.2,Issue.4,2017,pp:364-368.

[12] Niketa Gandhi, leisa J.Armstrong, "Application of data mining techniques for predicting rice crop yield in Semi-Arid climate zone of India",*IEEE International conference on technological Innovations in ICT for Agriculture and rural Development*,2017,pp:116-120.

[13] Manal Abdullah and Salma Al-Asmari,"Anemia types prediction based on data mining classification algorithms",*Communication, Management and Information Technology – Sampaio de Alencar (Ed.)*,2017,pp:615-621.

[14] K. Lakshmi, D.Iyajaz Ahmed, G. Siva Kumar," A Smart Clinical Decision Support System to Predict diabetes Disease Using Classification Techniques", *IJSRSET*,vol.4,Issue.1,2018,pp: 1520-1522.

[15] Himansu Das, BighnarajNaik and H. S. Behera,"Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", *Springer Nature Singapore Pte Ltd*, 2018, pp: 539-549.

[16] Miss. N. Vijayalakshmi, Miss. T. Jenifer,"An Analysis of Risk Factors for Diabetes Using Data Mining Approach",*International Journal of Computer Science and Mobile Computing*, Vol.6, Issue.7, 2017, pp:166 – 172.

[17] B. Ahmed Mohamed Ahmed, AhmetRizanerc, Ali HakanUlusoyc," Using data mining to predict instructor performance",*12th International Conference on Application of Fuzzy Systems and Soft Computing*, 2016, pp: 137 – 142.

[18] C. Kalaiselvi and G. M. Nasira, "Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques", *Indian Journal of Science and Technology***,** Vol: 8(14), 2015.

[19] Mr. A. Amol , Wghmode, Mr. DarpanSawant, Prof. Deven D. Ketkar, "Heart Disease Prediction Using Data mining Techniques", *International Journal of Engineering Technology Science and Research*,vol:4,Issue:10,2017,pp:366-369.

[20] Jay Gholap," Performance Tuning of J48 Algorithm for Prediction of Soil Fertility", *Department Of Computer Engineering, College Of Engineering, Pune, Maharashtra, India*, 2017.

[21] Miss.SnehalS.Dahikar, Dr.SandeepV.Rode**,"** Agricultural Crop Yield Prediction Using Artificial Neural Network Approach**",** *International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering*, Vol.2, Issue .1, 2014, pp: 683-686.