International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Diabets Dataset Classification and Analysis

R.V.S.S. Nagabhushana Rao[1], V.Munaiah[2], J. Prabhakara Naik[3], K.Vasu[4], G.Mokesh Rayalu[5]
[1]*Assistant professor, Department of Statistics, Vikrama Simhapuri University, Nellore*
[2]*Lecturer , Department of Statistics, PVKN Govt College, Chittoor*
[3]*Lecturer in Statistics, SSBN Degree & PG College(Autonomous), Anantapur*
[4]*Assistant Professor, Vidya Educational Institutions ,Yadamari , Chittoor*
[5]*Assistant Professor ,Department of Mathematics, School of Advanced sciences ,VIT, Vellore*

Abstract: Data mining is the technique which is used to identify hidden information from our datasets this naïve-Bayes statistical analysis produce the report for classification. Instead of naïve-Bayes we can utilize more efficient classification algorithms. XLMiner is designed efficiently to provide detailed analysis result for evaluating student levels in their examination and give an analytical report for early school leavings with the help of machine learning techniques. Problem view can be represented by the data mining preprocess techniques. Large complex datasets in excel are handles by XLMiner. WEKA gives detailed information and gives an analysis report for our datasets. Synopsis operator can be used for evaluating classification tasks. We are utilizing classification techniques to analyze different datasets. Here we are using open source tools XLMiner, WEKA. in these applications we are using naïve-Bayes algorithm to evaluate student performance.

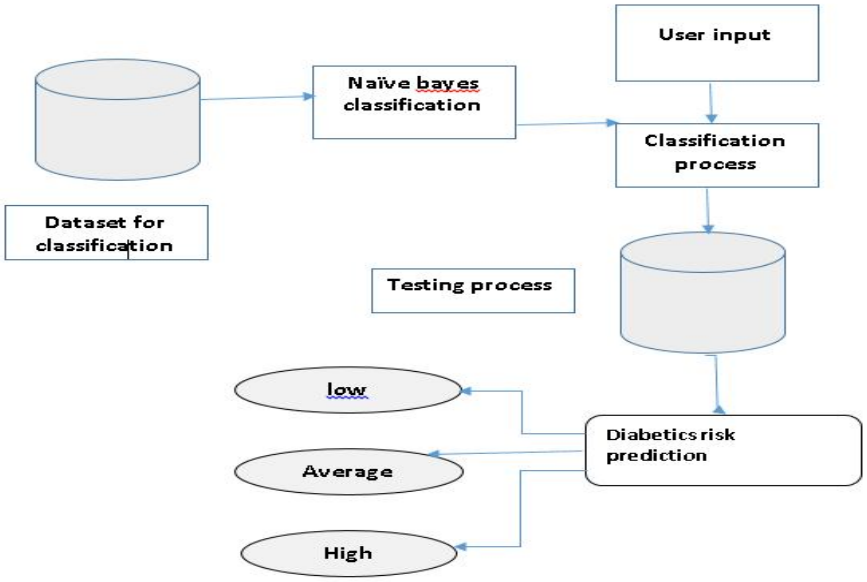
I. INTRODUCTION

Due to an increasing amount of data, we meet complexity to analyze data manually. Data plays a major role in every industries, organizations for decision making strategy. in this case we use data mining strategies to get comprehensive data from the available data. Classification includes the rules for partitioning given data. it is the process of identifying information from large amounts of data which is in data warehouse and database. Here we perform
comparison between classification using different classification tools like XLMiner, WEKA, Rapid Miner. Correctly classified data is used to evaluate the dataset performance. Rapid Miner is an open source environment for machine learning and preprocessing techniques. Complex design will be created for nested operator which is used for data input and output for different formats. Which joins huge amount of problems. to describe discovery process it uses XML.

A. Dataset description:

DATASET: Diabetics				
Attributes	4			
instance variables	Pregnancies	Blood Pressure	Skin Thickness	Age
Output value	Outcome			

II. LITERATURE SURVEY

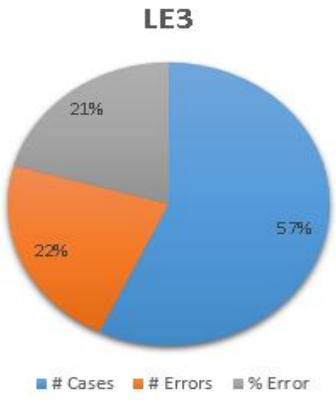


Dataset is taken for perform naive Bayes classification .user input will be given to perform process .this classified information will be stored in database. Then it will be applied for diabetics risk prediction. And testing process also performed. And this prediction list will be classified as low average and high predictive analysis. Wetake one dataset for performing classification using xlminer. XL Miner is an add-ins for Microsoft excel. it has different evaluation techniques to analyze datasets. it partitions the dataset into two types one for training the dataset and the other for testing the dataset. Xlminer-we can load excel sheet in xlminer. it offers statistical methods and also machine learning techniques. in this xlminer neural network has highest classification rates. Naïve Bayes classifier performs classification and estimate probabilities. it does not require lots of observations for the instance variables. Difficulty in this classifier is you cannot upload dataset which contains millions of observations.

II. RESULT ANALYSIS

Error Report			
Class	# Cases	# Errors	% Error
LE3	256	100	92.43428
GT3	399	51	1.895576
Overall	282	121	23.2289

Here we get large number of error report occurences .it takes more time for process.



Correctly Classified Instances	586	76.3021%
Incorrectly Classified Instances	182	23.6979%
Kappa statistic	0.4664	
Mean absolute error	0.2841	
Root mean squared error	0.4168	
Relative absolute error	62.5028	
Root relative squared error	87.4349	
Total Number of Instances	768	

By analyzing this we have errors in attributes LE3 is 100.and attributes GT3 is 51 and overall attributes are 121.by using xlminer tool we get more error report.it is not suitable for large datasets. For large number of attributes we have to separate the dataset .and analyze the two different dataset. Here we have 92% error report it consist confusion matrix value, predicted class and actual class report. We need only independent instances and attributes. it will perform analysis with comprehensive datasets. Researching the analysis and comparing this result with diabetic patients record.

III. PROPOSED SYSTEM

A. Preprocessing

Missing values are identified filters which is a major function of weka tool. Preprocessing plays a major role for removing noises and incompatible text and other values. it is very rare to get cleaned data and also handling noises in that dataset. If we develop techniques for machine learning we can achieve very better performance data. Diabetes dataset consist some attributes like blood pressure level and body health level. This two attributes are specified manually. We can handle this type of noises in weka using numeric cleaner filter. You can identify and remove missing values by doing following steps. If dataset contains more number of mishandled data open your weka console application in console open explorer. And then open your csv file location directory select file. And choose numeric cleaner under filter. it is an unsupervised learning. You can also handle missing values by following simplest way you can permanently delete or remove this values by remove with values in filter. You can select filters under filter menu and open.it is necessary for selecting instance variables for a model. it includes more number of records.

WEKA-freely available tool supports many preprocessing techniques like regression, classification, clustering and visualization techniques .it also performs machine learning techniques. We can process the datasets which is in .ARFF file format. And other types of datasets are also converted into .ARFF files using WEKA file format converters. This files consist two separate segments. First segment have the table name, instances, attributes and their types. After validating this all it will load the dataset and perform specified classification or other preprocessing techniques. in weka system defined techniques are used for perform classification .here we are using user defined classifications and clustering techniques. Weka-It is a knowledge learning tool .it identifies the hidden data from datasets. Using Weka we can perform many analytical process. Summary report cab be in visualization form.it is used to perform many data visualization techniques like classification, clustering, regression. You can use nominal dataset, numeric dataset, and alpha numeric dataet. it support .arff file format. You can also load csv file from weka explorer. We get following analysis report .WEKA it contains tool for all preprocessing technique. For research and academic purpose WEKA plays a major

role in data discrediting. Data mining in diabetics prediction performance evaluation makes data into comprehensive information. Decision making process done using this detailed data. in this dataset it does not consist any missing values. Min-max normalization performs data preprocessing technique, if it consist any missing value. After identification of missing values it destroy the missing values. it maps the new value instead of previous one.

B. Result Analysis

correctly classified instances are obtained as 30% and it also considered as sample accuracy .and incorrectly classified instances 27%.and kappa statistic report is 1%.we get mean absolute error is 43%.Select classify button under the xlminer add ins which will give detailed classification algorithms. Then identify the missing values by preprocessing it. Choose any classification algorithm and visualize it. finally it creates an report analysis which contains correctly classified values ,confusion matrix and error report .Confusion matrix contains actual and predicted class . Error report contains accuracy, precision and recall values. .time complexity is less and at the same time we can load large amount of dataset .it is very efficient to handle billions of records .replaced by numeric value. it gives by analyzing the performance of different

IV. CONCLUSION

Weka is best for analyzing large datasets with losing any data. by using weka there is no loss in dataset and missing values are tools which perform naïve bayes classification we conclude that weka tool is the best for millions of records.

REFERENCES

[1] ”Using methodologies of data miningin diabetics diagnosis and treatment” Mai Shouman, Tim Turner, Rob Stocker,May-August 2010.Egypt Conference on Electronics, Communications and Computers.vol-3.
[2] “Analysis of classification techniques and datamining over diabetics disease” andAaditya Sunder, P. PushpaLatha -vol-2,May-June 2012.
[3] “Data Mining Techniques and Concepts”, N. Aaditya Sunder, P. PushpaLatha Morgan Kaufmann Publishers, 2006.vol-3.May-June 2010.
[4] Naive Bayesian Classification techniques and Approach in diabetics Applications International Journal R. Bhuvaneswari and K. Kalaiselvi,vol-4,June-July 2013.
[5] “Using AssociativeClassifiers for Naïve Bayes Classification Analysis in Health CareData Mining”, Sunita Soni, O.P.Vyas, International Journal of ComputerApplication .Volume 4–5 No.10, July 2010, pages 33-34.
[6] “Prediction of Diabetic Dataset for Portuguese schools based on Their best Performance Using Decision Tree in Weka Environment ” Jaimin N. Undavia Prashant M. Dolia,Ph.D Nikhil P. Shah .
[7] “Educational Data mining for Prediction of Student Performance Using Classification and Clustering Algorithms” ,International Journal of Computer Science and Information Technologies, M. Durairaj, C. Vijitha , Vol. 6 (4)pages 23-24 , 2014.
[8] “Data Mining Techniques for performing Intellectual Performance Analysis of Students”Volume 3, Special Issue 3, March 2014-2015.
[9] “ Implication of Classification Techniques and Analysis in Predicting Student performance”Anupama Kumar
[10] ,“Students Performance Prediction System using Multiagent Data mining

.