ANALYZING STUDENT PERFORMANCE IN BOYS HIGHER SECONDARY SCHOOL USING DATA MINING TECHNIQUES

E. Sivagamasundari[1], Dr.N.R.Ananthanarayanan[2]

[1]M.Phil Research Scholar, Dept of CSA, SCSVMV,Enathur, Kanchipuram. TamilNadu, India –631561

[2]Associate Professor, Dept of CSA, SCSVMV,Enathur, Kanchipuram. TamilNadu, India –631561

ABSTRACT

In recent years it is been observed that analysis and evaluation of students‟ performance and retaining the standard of education has become an important aspect in all the educational institutions. This paper is analyzing and evaluating the school students‟ performance using data mining classification algorithms with the help of weka tool. Data mining tool is globally accepted as a decision making tool to ease better resource use in terms of student's performance. Some of the classification algorithms available are Random Forest, J48, Multilayer Perceptron, Decision Table and IB1 are being implemented in this paper. The results of such classification model deals with confusion matrices, accuracy level and also execution time.

Keywords:-Students Dataset, algorithms, Data mining, classification, pre-processing, Weka**.**

I. INTRODUCTION

Data mining is used by companies with a strong consumer focus - financial, retail, communication, and marketing organizations. It helps companies to understand the relationships among "internal" factors like product positioning, price or staff skills. Data mining is being used in many areas to analyze huge amount of data. In general, data mining also called as knowledge discovery is the process of analyzing data taken from different datasets and summarize them into useful information. Data mining software is an analytical tool to analyze data. It enables users to analyze data from many various angles or dimensions, and summarize the relationship that was identified. In technical terms, data mining is the process of finding patterns or correlations in large relational databases.

Educational data mining (EDM) that is recently discovered with the purpose of machine learning, data mining and statistics that is to be applied on the information generated from educational settings e.g., universities and intelligent tutoring systems. Online Examination is one of such application.

This site helps institutes to register and host online exams. This site was implemented to eliminate the flaws in manual system of conducting exams. All that the user needs to do in order to use this application is to register on the site and then enter the exam details and the lists of the students who will appear for the examination. This approach eliminates the physical location set up for examination and also displays the results immediately thereby reducing effort and time of both the students and the institutions. The growth of Information and Communication Technology has made a huge mark around the world. It helped to achieve a professional level in the field of education with the advantages of lower cost and time. This study analyses student's performance using data mining technique like decision tree algorithm, classification with the help of a classifier model based on the responses of students to courses evaluation questions.

II. RELATED WORK

Han and Kamber [3] defines data mining software that helps the users to analyze data from different dimensions, categorize and summarize the relationships that are identified in the mining process. to both the patient and the doctor by lab personnel using this functionality.

Brijesh Kumar Baradwaj and Saurabh Pal [1] defines the main aim of higher education institutions is to deliver quality education to students. One way to achieve this goal is by discovering knowledge for prediction regarding enrolment of students in a particular course. Also to detect abnormal values in the result sheets of the students, prediction about students' performance and so on. The classification process is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here.

Alaa El-Halees [4] implemented educational data mining ideas for developing methods in discovering knowledge from data derived from educational environment. The author used educational data mining to analyze the learning behavior of students. Student's data was collected from the Database course. Once data is preprocessed, we applied data mining techniques to identify the classification, association, and clustering and outlier detection rules. In each of these four tasks, we extracted knowledge that describes students' behavior.

Mohammed M. Abu Tair and Alaa M. El-Halees [5] implemented educational data mining ideas for developing methods in discovering knowledge from data derived from educational environment. The author used educational data mining to analyze the performance of students in order to overcome the problem of low grades of graduate students and try to extract useful knowledge from graduate students. Data for the study was collected from the college of Science and Technology. The data includes fifteen years period [1993-2007]. Once data is preprocessed, we applied data mining techniques to identify the classification, association, and clustering and outlier detection rules. In each of these four tasks, we extracted knowledge and described its importance in educational domain.

G. N. Pandey, SonaliAgarwal, and M. D. Tiwari [6] explain that educational organizations are important part of our society and play a vital role for growth and development of our nation. Data Mining is an emerging technique to efficiently learn with historical data and use the accquired knowledge to predict the future behavior of concern areas.

Monika Goyal and RajanVohra [7] implemented data mining techniques to enhance the efficiency of higher education institution. Data mining techniques like decision tree and association, clustering are applied to higher education processes so as to improve students' performance, selection of courses, their life cycle management, to measure their retention rate and the grant fund management of an institution.

BrijeshBharadwaj, Surjeet Kumar Yadav, and Saurabh Pal [11] applied decision tree classifiers in order to find the best classifier for retention data to predict the student's drop-out possibility.

Brijesh Kumar Baradwaj and Saurabh Pal [12] implement the classification task on student database in order to predict the students division on the basis of previous database

III. OBJECTIVE

This work implements data mining process in a student's database using classification data mining techniques. The result generated after the analysis of data mining techniques on student's data base is useful for the executives for training & placement department of engineering colleges. This work categorizes student's performance in their academic qualifications. In future this study will be helpful for industries and institutions. We can also generate information using other data mining techniques like Predication and Association rules, clustering, etc on different eligibility criteria of industry recruitment for students

METHODOLOGY

DATA PREPROCESSING

Data have quality if it serves the required purpose. There are several factors comprising data quality that includes completeness, accuracy, consistency, believability, timeliness and interpretability.

Let's assume that you are a manager at All Electronics and you have are in charge to analyze company's data. So you fetch company's database and select the required attributes (e.g., item, price, and units sold). You notice that several attributes for various tuples hold no record value. You also notice that the information about each item purchased was advertised as on sale is missing or not recorded. There are also unusual values, errors and inconsistencies in the data recorded for some transactions. So the data you wish to analyze using data mining techniques are incomplete, inaccurate or noisy and inconsistent. Incomplete, inaccurate and inconsistent data are common properties of large databases. There are various reasons for inaccurate data. It could be the cause of a faulty instrument, human or computer errors. Errors during data transmission are also possible. There can be limited buffer size for coordinating synchronized data transfer and consumption. Duplicate tuples also require data cleaning. Timeliness is also a factor that affects data quality. Let's assume that you are overseeing the distribution of monthly sales bonuses to the top sales representatives at AllElectronics. Many sales representatives, however have failed to submit their sales records on time. There are also a number of adjustments and corrections that flow in after the month's end. The fact that the month-end data is not being updated on a timely basis has a negative impact on the entire data quality.

Two other factors that also contribute to data quality are interpretability and believability. Believability means the trust on the data and interpretability means the ease to understand the data. Suppose that a database, at one point, had several errors, all of which have since been corrected. The past errors might have caused several problems for sales department users, and hence they no longer trust the data. The data also use many accounting codes, that is not understood by the sales department. Though the database is now complete, accurate, consistent, and timely, sales department users may regard it as of low quality because of the poor interpretability and believability.

IV. CLASSIFICATION

Classification is being implemented in order to process the data and classify them into predefined categorical class labels. Classification consists of two step process namely training and testing. In the initial stage a model is designed by analyzing the data tuple from the given training data. Now for each tuple preset in the training data, the worth of class label attribute is being understood. The defined classification rule is then applied on training data in order to form the model. In the next step of classification, the test data is engaged to observe the accuracy of the model. If the accuracy of the model meets the expectation then the model can be used to classify the unknown data tuples. The fundamental techniques to perform classification are neural networks, decision tree classifier, Lazy based classifier and rule based classifier

CLASSIFICATION ALGORITHMS

This research paper implements the J48, NAVE BAYES, RANDOM FOREST, BAYES NET, LOGISTICS and REPT TREE.

DATA CLEANING AS A PROCESS

Missing values, inconsistencies and noise contributes to inaccurate data in the data set. Along with smoothing data and missing data, cleaning data is also a task to be performed.

Data cleaning deals with discrepancy detection. Discrepancies can happen due to poorly designed data entry forms that comprises of many optional fields, deliberate errors (e.g., respondents not wanting to provide the information), human error in data entry, and data decay (e.g., outdated addresses). Discrepancies also happen due to inconsistent data representations and inconsistent use of codes. Other factors that contribute for discrepancies are system errors and errors in instrumentation devices that record data. Inconsistencies can occur due to data integration as well.

In order to handle discrepancy in data, use any knowledge that you may already possess about the data. Such knowledge is referred as metadata. You can study about the data types, attributes, and domain and identify the acceptable values about the dataset like the median, mean, and mode values. You can also find if the data is symmetric or skewed. After finding as much as information about the data you can write your own scripts and/or use some of the tools that we discuss further later. Now you can identify noise, unusual values and outliers that need investigation.

The data must be examined for consecutive rules, unique rules and null rules. A unique rule defines that each value of the given attribute is different from all other values for that particular attribute. A consecutive rule states that there must be no missing values between the lowest and highest values for that particular attribute, and also all values must be unique (e.g., as in check numbers). A null rule defines that the use of question marks, blanks, special characters, or other strings may indicate the null condition (e.g., where a value for a given attribute is not available), and how such values must be handled.

There are several different commercial tools that can help in handling discrepancy detection step. Data scrubbing tools uses simple domain knowledge (e.g., knowledge of postal addresses and spell-checking) to identify errors and make the necessary corrections in the data.

These tools depend on the fuzzy and parsing matching techniques while cleaning data from multiple sources.Data auditing tools helps to find discrepancies by observing the data to discover relationships and rules, and can also detect data that violates such conditions.

Data migration tools helps to perform simple transformations like replace the string “gender” by “sex.” ETL (extraction/transformation/loading) tools allow users to specify a particular transform using a graphical user interface (GUI). WE can also write custom scripts for this step of the data cleaning process.

There are new approaches to data cleaning where interactivity is being emphasized. For example, Potter’s Wheel is a data cleaning tool that integrates transformation and discrepancy detection. Another approach in order to increase interactivity in data cleaning is the declarative languages for the specification of data transformation operators. This approach focuses on defining powerful extensions to SQL and algorithms that helps users to express data cleaning specifications well. As and when we discover more about the data, it is important to keep updating the metadata to reflect this knowledge. This will improve the speed of data cleaning on future versions of the same data store.

V. DATA INTEGRATION

Data mining can be efficiently performed if data integration is done. Data integration is merging of data from multiple databases or data stores. Careful integration can actually help in reducing and avoiding inconsistencies and redundancies in the data set. This can help to increase the accuracy and maximize the speed of the data mining process.

VI. WEKA

Weka is a collection or a set of machine learning algorithms that are particularly meant for data mining tasks. These algorithms can be applied directly to a dataset or also from your own Java code. Weka comprises of tools used for data pre-processing, regression, classification, clustering, visualization and other association rules. It is the best technique to develop new machine learning schemes.

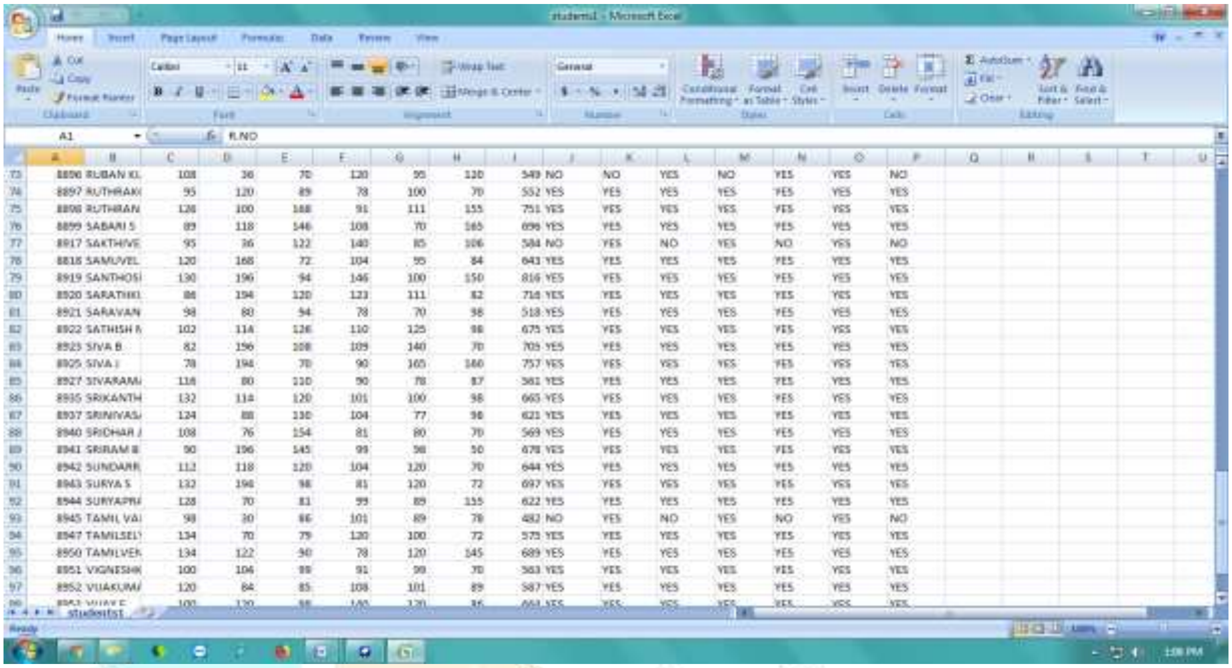
Weka is applied in several fields like education, research and applications. It comprises of a comprehensive set of learning algorithms, data pre-processing tools, evaluation methods, environment for comparing learning algorithms and graphical user interfaces (incl. data visualization).

Weka is open source software that was issued under the GNU General Public License. "WEKA" expands as Waikato Environment for Knowledge Analysis that was developed at the University of Waikato located in New Zealand. WEKA has become a collection ofmachine learning algorithms to solve real-world data mining concerns and problems. It is written in Java and can be executed on almost all platforms.

Weka is simple and easy to use and can be applied at various different levels. The WEKA class library can be used from your own Java program, and can also be implemented on new machine learning algorithms. The three major implementation schemes involved in WEKA are: (1) Implemented schemes for classification. (2) Implemented schemes for numeric prediction. (3) Implemented "metaschemes”.

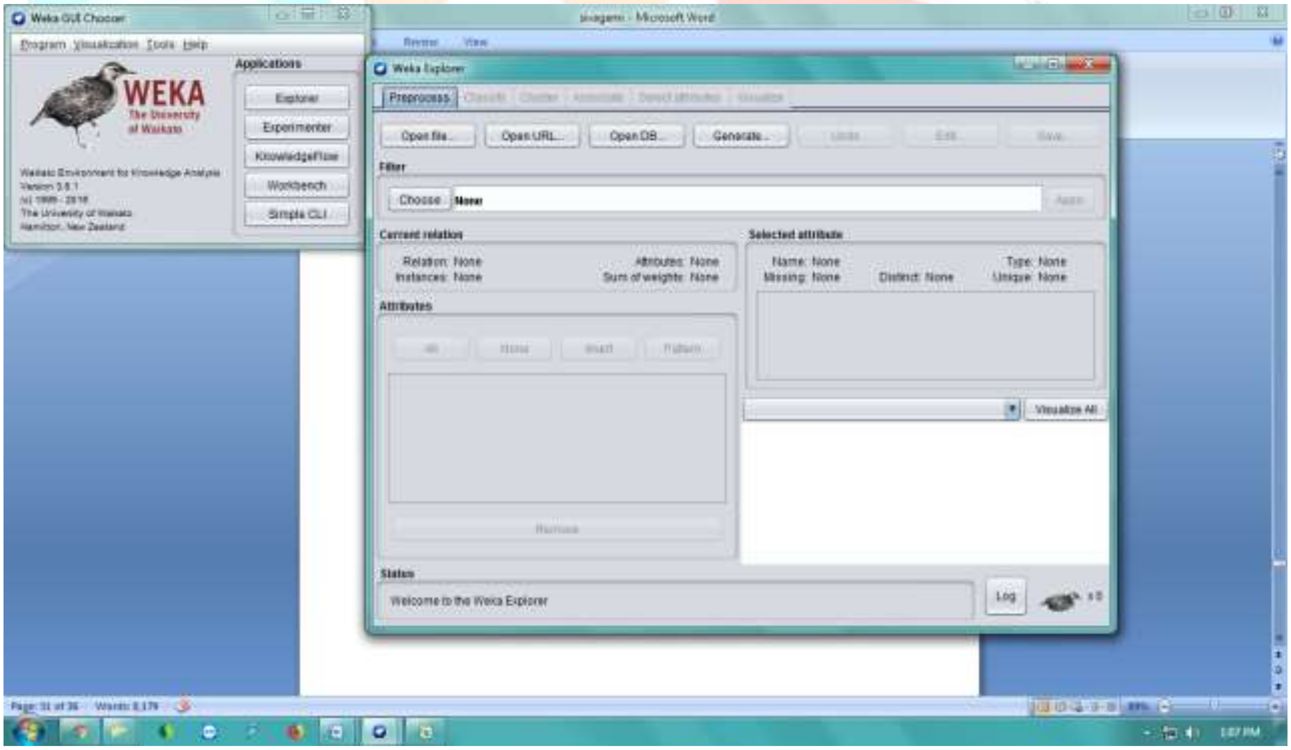
Apart from the actual learning schemes, WEKA also consists of a large variety of tools that are used for pre-processing datasets, so that the focus is kept undisturbed on your algorithm without worrying much on implementing filtering algorithm and providing code to evaluate the results and so on.

VII. RESULT



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
73	8896	RUBAN KL	108	36	70	120	95	120	549	NO	NO	YES	NO	YES	YES	NO
74	8897	RUTHRAKI	95	120	89	78	100	70	552	YES	YES	YES	YES	YES	YES	YES
75	8898	RUTHRAN	128	100	188	91	111	155	751	YES	YES	YES	YES	YES	YES	YES
76	8899	SABARI S	89	118	146	108	70	165	696	YES	YES	YES	YES	YES	YES	YES
77	8917	SAKTHIVE	95	36	122	140	85	106	584	NO	YES	NO	YES	NO	YES	NO
78	8818	SAMUVEL	120	168	72	104	95	84	643	YES	YES	YES	YES	YES	YES	YES
79	8919	SANTHOSI	130	196	94	146	100	150	816	YES	YES	YES	YES	YES	YES	YES
80	8920	SARATHKI	88	194	120	123	111	82	718	YES	YES	YES	YES	YES	YES	YES
81	8921	SARAVAN	98	80	94	78	70	98	518	YES	YES	YES	YES	YES	YES	YES
82	8922	SATHISH K	102	114	126	110	125	98	675	YES	YES	YES	YES	YES	YES	YES
83	8923	SIVA B	82	196	108	109	140	70	705	YES	YES	YES	YES	YES	YES	YES
84	8925	SIVA J	78	194	70	90	165	160	757	YES	YES	YES	YES	YES	YES	YES
85	8927	SIVARAM	118	80	110	90	78	87	581	YES	YES	YES	YES	YES	YES	YES
86	8935	SRIKANTH	132	114	120	101	100	98	665	YES	YES	YES	YES	YES	YES	YES
87	8937	SRINIVAS	124	88	130	104	77	98	621	YES	YES	YES	YES	YES	YES	YES
88	8940	SRIDHAR	108	76	154	81	80	70	569	YES	YES	YES	YES	YES	YES	YES
89	8941	SRIRAM B	90	196	145	99	98	50	678	YES	YES	YES	YES	YES	YES	YES
90	8942	SUNDARR	112	118	120	104	120	70	644	YES	YES	YES	YES	YES	YES	YES
91	8943	SURYA S	132	198	98	81	120	72	697	YES	YES	YES	YES	YES	YES	YES
92	8944	SURYAPR	128	70	81	99	89	155	622	YES	YES	YES	YES	YES	YES	YES
93	8945	TAMIL VA	98	30	86	101	89	78	482	NO	YES	NO	YES	NO	YES	NO
94	8947	TAMILSEL	134	70	79	120	100	72	575	YES	YES	YES	YES	YES	YES	YES
95	8950	TAMILVEK	134	122	90	78	120	145	689	YES	YES	YES	YES	YES	YES	YES
96	8951	VIGNESHK	100	104	88	91	98	70	583	YES	YES	YES	YES	YES	YES	YES
97	8952	VIJAKUM	120	84	85	108	101	89	587	YES	YES	YES	YES	YES	YES	YES

Fig.1.Student Data set in CSV Format



Waikato Environment for Knowledge Analysis
Version 3.8.1
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand

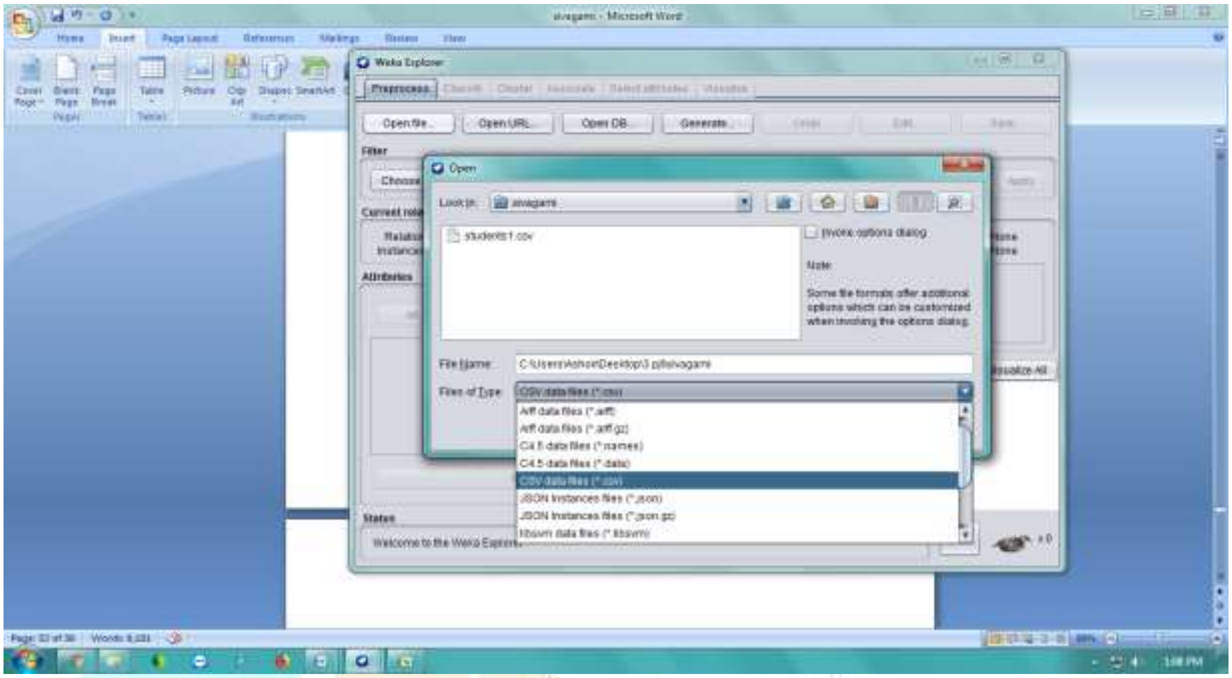
Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

Fig.2.Weka Explorer Menu



Open

Look In: sivagami

students1.csv

□ Invoke options dialog

Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: C:\Users\Ashok\Desktop\3 p\sivagami

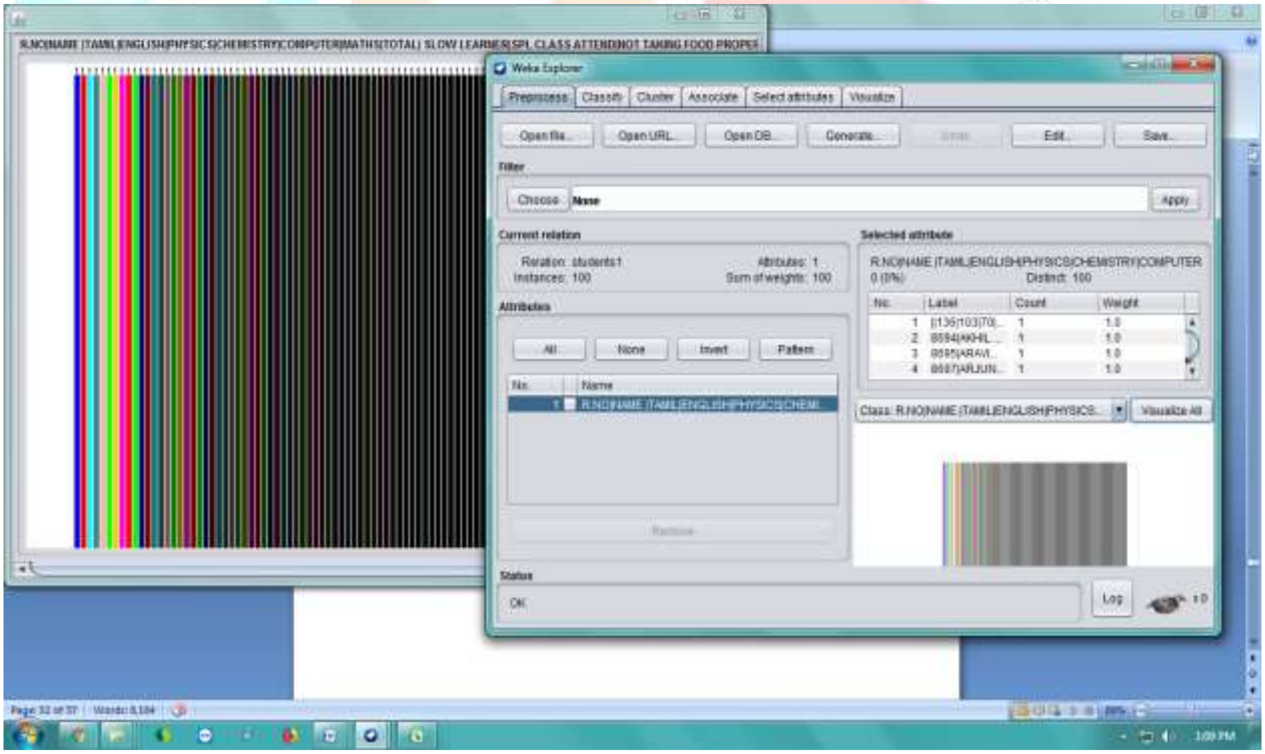
Files of Type: CSV data files (*.csv)

Arff data files (*.arff)
Arff data files (*.arff.gz)
C4.5 data files (*.names)
C4.5 data files (*.data)
CSV data files (*.csv)
JSON Instances files (*.json)
JSON Instances files (*.json.gz)
libsvm data files (*.libsvm)

Status

Welcome to the Weka Explorer

Fig.3.Student Data set Selection



R.NO|NAME |TAMIL|ENGLISH|PHYSICS|CHEMISTRY|COMPUTER|MATHS|TOTAL| SLOW LEARNER|SPL CLASS ATTENDING|NOT TAKING FOOD PROPE

Filter: Choose None

Current relation

Relation: students1 Attributes: 1
Instances: 100 Sum of weights: 100

Selected attribute

R.NO|NAME |TAMIL|ENGLISH|PHYSICS|CHEMISTRY|COMPUTER
0 (0%) Distinct: 100

No.	Label	Count	Weight
1	|1136|103|70|...	1	1.0
2	8694|AKHIL...	1	1.0
3	8695|ARAVI...	1	1.0
4	8697|ARJUN...	1	1.0

Attributes

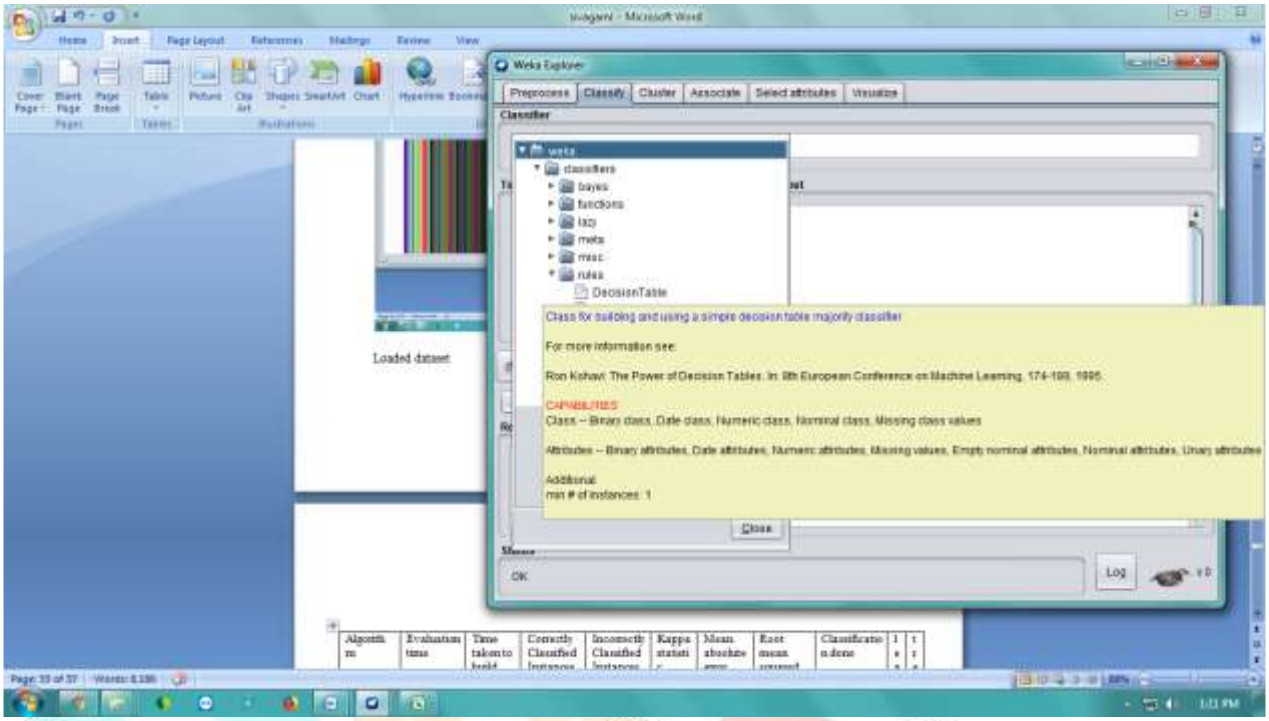
No.	Name
1	R.NO|NAME |TAMIL|ENGLISH|PHYSICS|CHEMI

Class: R.NO|NAME |TAMIL|ENGLISH|PHYSICS...

Status

OK

Fig.4.Loaded Student dataset



Class for building and using a simple decision table majority classifier

For more information see:

Ron Kohavi: The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189, 1995.

CAPABILITIES
Class -- Binary class, Date class, Numeric class, Nominal class, Missing class values

Attributes -- Binary attributes, Date attributes, Numeric attributes, Missing values, Empty nominal attributes, Nominal attributes, Unary attributes

Additional
min # of instances: 1

Fig. 5.Algorithm selection Process

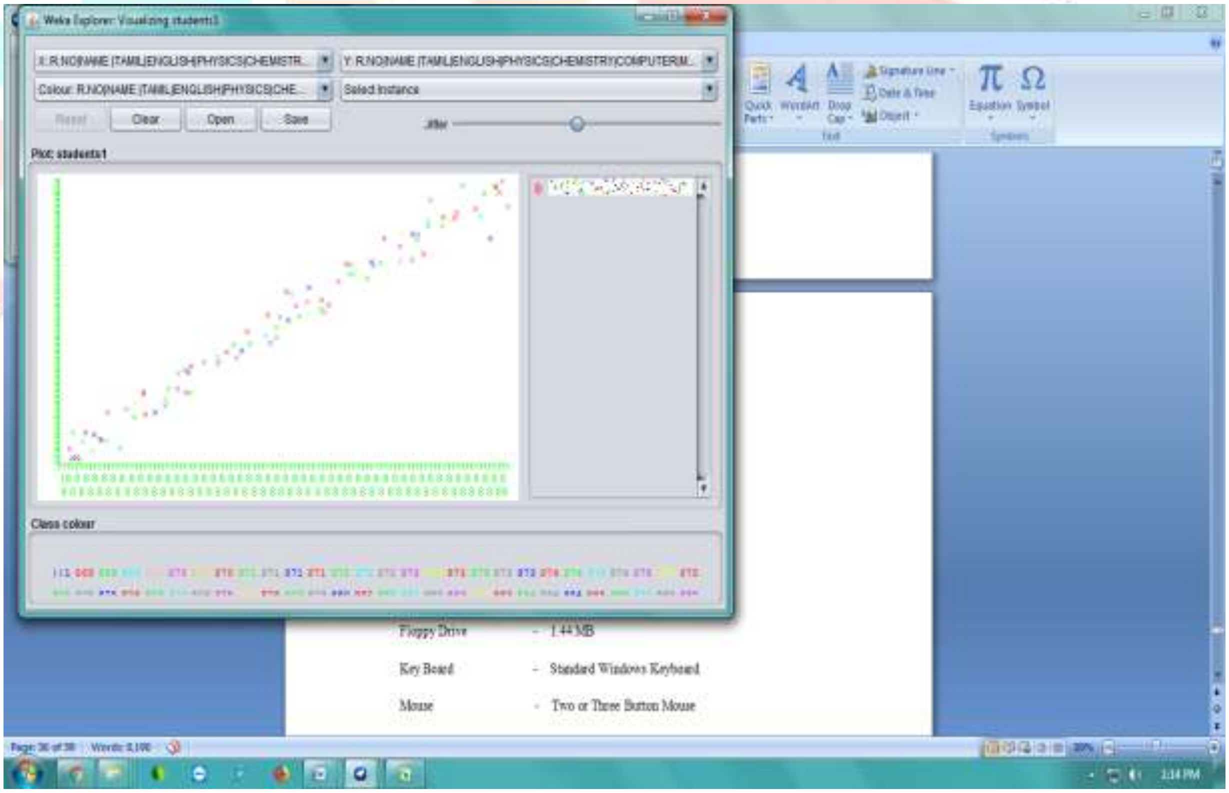
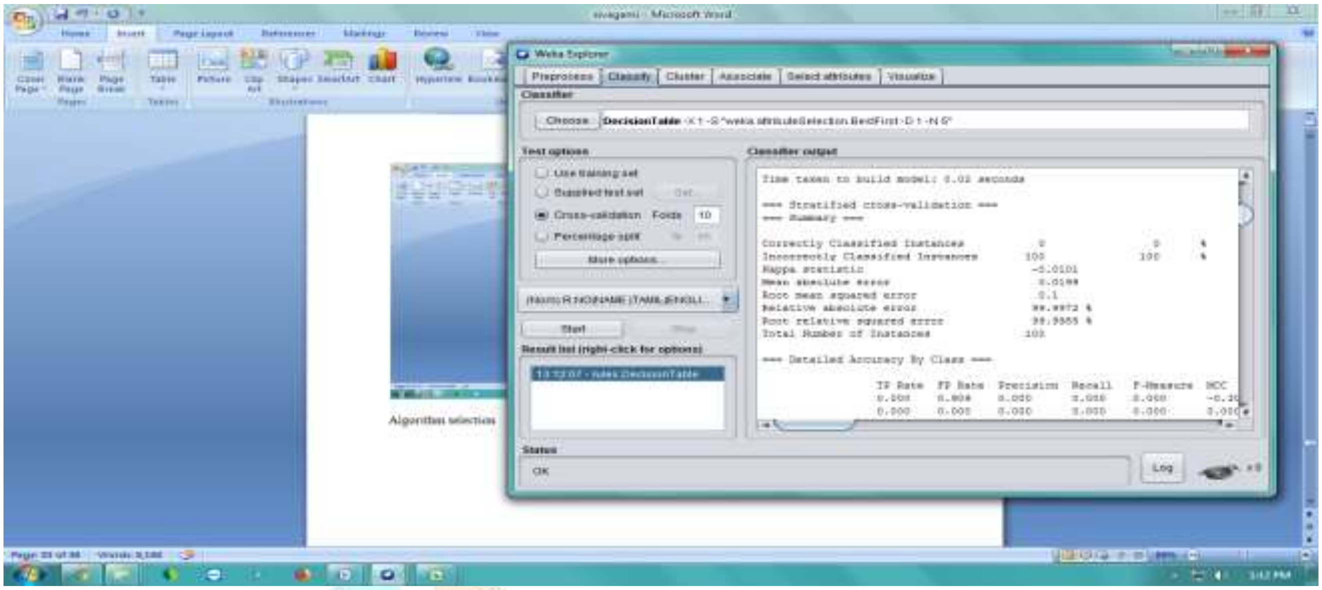

Weka Explorer: Visualizing students1
X: R.NO|NAME |TAMIL|ENGLISH|PHYSICS|CHEMISTR...
Y: R.NO|NAME |TAMIL|ENGLISH|PHYSICS|CHEMISTRY|COMPUTER|M...
Colour: R.NO|NAME |TAMIL|ENGLISH|PHYSICS|CHE...
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: students1
Class colour
Floppy Drive - 1.44 MB
Key Board - Standard Windows Keyboard
Mouse - Two or Three Button Mouse

Fig.6.Classification of Students Dataset



DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           0               0      %
Incorrectly Classified Instances       100             100      %
Kappa statistic                         -0.0101
Mean absolute error                      0.0198
Root mean squared error                  0.1
Relative absolute error                 99.9972 %
Root relative squared error             99.9985 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   ROC
                 0.000     0.808     0.000       0.000    0.000       -0.10
                 0.000     0.000     0.000       0.000    0.000       0.000
```

Fig.7.Visualizing result of Students Dataset

Table-1: Comparison of Algorithms and its Classification Accuracy applying Students Dataset

Algorithm	Evaluation time	Time taken to build model	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Classification Accuracy
NAVE BAYES	0.01	0 s	99%	1 %	0%	0.0198	0.0995	100%
BAYES NET	0.03	0 s	99%	1%	0%	0.0198	0.0995	99%
J48	0.01	0.01s	97%	3%	0%	0.0198	0.0995	99%
RANDOM FOREST	0.4	0.23s	97%	3%	-0.0101%	0.0198	100.0016	98%
RANDOM TREE	0.23	0.23s	100%	0%	-0.0101%	0.0198	0.1	97%
REPT TREES	0	0s	97%	3%	-0.0101	0.0199	0.1	100%

Conclusion

There are students graduating from colleges and universities every year and the feedback from the students is used to apply the data mining technique. Student Dataset is applied to different algorithms which shows the classification accuracy which shows 100% and 97% which shows that the factors such as lack of proper food, not attending special classes and always spending time in sports which affects the academic performance of students and also in order to take the necessary steps like motivating the students, giving proper guidance for higher education, giving gifts for students who attend the classes properly and also various aspects of learning process would improve the quality of education in future.

REFERENCES

[1] Brijesh Kumar Baradwaj, Saurabh Pal, Data mining: machine learning, statistics, and databases, 1996.

[2] Nikhil Rajadhyax, RudreshShirwaikar, Data Mining on Educational Domain, 2012.

[3] JiaweiHan ,MichelineKamber, Data Mining: Concepts and Techniques, 2nd edition, 2006.

[4] Alaa El-Halees, Mining Students Data to Analyze Learning Behavior: A Case Study, 2008.

[5] Mohammed M. Abu Tair, Alaa M. El-Halees, Mining Educational Data to Improve Students' Performance: A Case Study, 2012.

.

[6] SonaliAgarwal, G. N. Pandey, and M. D. Tiwari, Data Mining in Education: Data Classification and Decision Tree Approach, 2012.

[7] Monika Goyal ,Rajan Vohra2, Applications of Data Mining in Higher Education, 2012.

[8] P. Ajith, M.S.S.Sai, B. Tejaswi, Evaluation of Student Performance: An Outlier Detection Perspective, 2013.

[9] Varun Kumar, AnupamaChadha, An Empirical Study of the Applications of Data Mining Techniques in Higher Education, 2011.

[10] Hongjie Sun, Research on Student Learning Result System based on Data Mining, 2010.

[11] Surjeet Kumar Yadav, BrijeshBharadwaj, and Saurabh Pal, Mining Education Data to Predict Student's Retention: A comparative Study, 2012.

[12] Brijesh Kumar Baradwaj, Saurabh Pal, Mining Educational Data to Analyze Students‟ Performance, 2011.

[13] K.ShanmugaPriya, A.V.Senthil Kumar, Improving the Student's Performance Using Educational Data Mining 2013.