ANALYSIS OF STUDENT PERFORMANCE IN GOVT GIRLS HIGHER SECONDARY SCHOOL USING DATA MINING TECHNIQUES

B. M. Vijayalakshmi[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

Data mining methodology is considered as an important contribution in researches related to extraction of the hidden knowledge and information from a given set of data. Data mining is interrelated with developing and enhancing methods to discover knowledge from data. This paper deals with the concept of categorizing students based on grade, in order to improve their education studies and to measure their academic performances. This paper implements the educational data mining techniques to develop 'students' performance by applying appropriate algorithms and see the overall accuracy of the students.

Keywords: - Students Data, academic , category, pre-processing, classification

INTRODUCTION

We are aware of the remarkable improvement in data mining and its application in the educational sector. The modern engineering called the educational data mining is to improve the methods involved in extraction of knowledge from data that is derived from the educational sector. Data mining is a technique to sort data and extract the hidden patterns from huge databases. This concept is being applied in several fields like medicine, marketing, real estate, engineering, customer relationship management, web mining, etc. Educational data mining is a new concept developed to study the data in the educational sector. The data required for this method is available from the past used data that resides in the databases of educational institutes. Data can be also used from e-learning database systems that hold huge educational information. Several techniques like neural networks, decision trees, naive bayes, K-Nearest neighbour and many others are used for the proper implementation of the technique. In order to sort and classify the huge data cluster is used. Cluster analysis is a method to classify the raw data in a reasonable way and divide them into several groups.

DATA MINING TECHNIQUES

Classification: Classification is a popular approach to find a set of functions or models that can describe and also distinguish data classes or concepts. Unlike the classification model, the prediction model helps to determine the future outcome rather than studying the current behavior. Its output can be numeric or categorical value. It is a supervised learning since it determines the classes before examining the data.

Clustering: It is helpful to identify similar data items in a dataset. Clustering is the process of partitioning data items into classes with similar characteristics items.

Association analysis: This analysis is used to identify the relationships between the attributes and the items such as the presence of one pattern implies the presence of another pattern. Association Rule is a popular technique in the field of market basket analysis since all possible combinations of interesting product groupings are explored.

Decision Tree: This is similar to a flow-chart design where the rectangular boxes are named as node. Here each node represents a set of records that is derived from the original data set. Internal node is a node that has a child and leaf (terminal) node is nodes that don't have children. Root node is the topmost node in the tree structure. The decision tree is helpful in finding the best way to distinguish a class from another class. The five most commonly used algorithms for decision tree are: CART, ID3, CHAID, J48 and C4.5 algorithm.

II. RELATED WORK

Mustafa Agaoglu [1] proposed a research work in educational mining focusing on modeling student's performance instead of instructors' performance. The common tools available to observe instructors' performance is the course evaluation questionnaire based on students' perception. This study implements four different classification techniques namely: support vector machines, decision tree algorithms, artificial neural networks, and discriminant analysis. The instructor's performances are compared over a dataset that holds the responses of students to a real course evaluation questionnaire through precision, accuracy, recall, and specificity performance metrics. All the classifier models demonstrate comparably high classification performances, but among all C5.0 classifier is the best in terms of precision, accuracy and specificity. In addition, the author has also performed an analysis of the variable importance for each classifier model. Accordingly, it was discovered that many of the questions in the course appear to be irrelevant. Furthermore, the study shows that the instructors' success is completely based on the students' perception.

TriptiMishra,Dr. Dr. Sangeeta Gupta, Dharminder Kumar [2] implements different classification techniques in order to build a performance prediction model that is based on students' academic integration, social integration and other emotional skills that are not been considered so far. Two algorithms Random Tree and J48 (Implementation of C4.5) are being applied to the records of MCA students of colleges affiliated to Guru Gobind Singh Indraprastha University to predict the students third semester performance. The Random Tree approach is found to be more accurate in predicting performance when compared to the J48 algorithm.

Keno C. Piad, Melvin A. Ballera, MenchitaDumlao, Shaneth C. Ambat [3] predicts the employment of IT graduates with the help of nine variables. Firstly, various classification algorithms in data mining are tested by making logistic regression with accuracy of 78.4. Based on the logistic regression analysis, three academic variables get directly affected; IT_Professional, IT_Core and Gender identified as significant predictors for employability. The data was derived from the five year profiles of 515 students who are selected randomly at the placement office tracer study.

BipinBihariJayasingh [4] initiates a sample study considering a particular institution, in a given environment, for the particular batch and students. The sample data was collected from a classroom through questionnaire to inquiry based and deductive learning. The system is developed and is tested twice after teaching the content using inductive method and implemented with the help of attribute relevance, discriminant rules of class discrimination mining. The results obtained are visualized in bar charts and shows that the selected two batches of different years have different learning characteristics.

S. M. Merchán [5] presents and analyzes the experience of using different data mining methods and techniques on 932 various systems engineering students' data, from El Bosque University in Bogotá, Colombia. It's an iterative approach and learning process and the results are obtained in each of the process' iterations. All the results are evaluated regarding the results that are expected, the data's input and output characterization, what theory dictates and the pertinence of the model obtained with respect to prediction accuracy.

KonstantinaChrysafiadi and Maria Virvou [6] selects a novel approach of web-based education that performs individualized instruction on the domain of programming languages. This approach was implemented in an educational application module, called as Fuzzy Knowledge State Definer (FuzKSD). FuzKSD works on a dynamic identification and updation the student's knowledge level for all the concepts of the domain knowledge. The operation of FuzKSD is based on the Fuzzy Cognitive Maps (FCMs). FuzKSD uses fuzzy sets to represent the students' knowledge level as a subset of the domain knowledge.

OBJECTIVE

Educational institutions have an important role to play in the society in terms of growth and development of our nation. Predicting student's performance in the educational environment is also important. Student's academic Education performance lies on various factors like social, personal details, psychological etc. The educational databases hold a lot of useful information in order to predict a students' performance and other details. Data mining techniques are helpful in these sectors to classify the educational database. Educational data mining deals with discovering methods to attain knowledge from data. This work aims to create a trust model with data mining techniques in order to mine the required information and helps the education system to adopt this as a strategic management tool.

METHODOLOGY

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.

DATA PREPROCESSING

Here a major step is involved in data preprocessing such as data cleaning, data integration and reduction, and finally data transformation.

Data cleaning is to “clean” the data by filling all the missing values in the database, smoothing the noisy data, identifying or removing the outliers, and also resolving inconsistencies. If the data in the database is dirty then it can result in a poor data mining application. Dirty data can also lead to confusion for the mining procedure and results in unreliable output. All the mining techniques will involve a procedure to deal with incomplete or noisy data. Instead, they can also work on avoiding over fitting the data to the function that is being modeled. Hence a useful preprocessing step is to perform data cleaning routines.

Now considering your task with AllElectronics, say that you want to inculcate several sources of data in your analysis. In other words you want to integrate multiple databases, files or data cubes. There arises a problem because some attributes that represent the similar data have different names in different databases. This can lead to inconsistency and redundancy of data. For example, the attribute named customer identification can be referred as customer id in one data store and cust id in another database. Naming inconsistencies can also be present in attribute values. For example, the first name can be registered as “Bill”, “William” or “B.” in various databases. It can also happen that some attributes may be inferred from others (e.g., annual revenue). A huge amount of redundant data can cause confusion or slow down the knowledge discovery process. Therefore data cleaning along with removal of redundant data is a must. Data cleaning and data integration are the preprocessing steps to prepare data for a data warehouse.

Now let’s say that you chose to work a huge database so it is sure that it will slow down the mining process. Is there a way to reduce the size of the data set without altering the data? Data reduction is a technique to achieve this by reducing the size of the data set that is much smaller in volume and still capable of producing the same analytical results. Data reduction includes dimensionality and numerosity reduction. In dimensionality reduction, data encoding schemes are used to achieve a “compressed” version of the original data. Some data compression techniques are principal components analysis and wavelet transforms, attribute subset selection (e.g., removing irrelevant attributes), and attribute construction (e.g., where a small set of more useful attributes is derived from the original set).Whereas in numerosity reduction, data is replaced by alternative or smaller representations using parametric models (e.g., log-linear models or regression) or nonparametric models (e.g., clusters, histograms, sampling, or data aggregation).

In order to enhance your data set further techniques like distance based mining algorithm like nearest-neighbor classifiers, neural networks or clustering can be used. These techniques provide better results when the data to be computed are normalized to a smaller range like [0.0, 1.0]. If the data set contains attributes like annual salary and age. For sure annual salary attribute will take larger values than age.Therefore, when the attributes are left unnormalized, the distance measurements taken on annual salary will generally outweigh the distance measurements taken on age.Discretization and concept hierarchy generation can be applied in such instances to replace by ranges or higher conceptual levels. For example, raw values for age may be replaced by higher-level concepts, like adult, youth, or senior.

Discretization and concept hierarchy generation allows data mining at multiple abstraction levels. Normalization, data discretization, and concept hierarchy generation are part of data transformation. These data transformation operations are to enhance data set further so as to achieve a successful mining process.

MISSING VALUES

- ✓ Let's say that you have to analyze All Electronics sales and customer data. You find that many tuples are not filled or there is no recorded value such as customer income. So in such instances what can you do about filling in the missing values for this attribute? Let's see the below options.
- ✓ Ignore the tuple: This is performed when a particular class label is missing (with the assumption the mining task involves classification). This approach is not so effective, unless the tuple contains of numerous attributes that has missing values. It performs poor when the percentage of missing values per attribute varies drastically. By ignoring the tuple, we also not use the remaining attributes' values in the tuple. Such data might have been useful to perform the task at hand.
- ✓ To fill the missing value manually: This approach is actually time consuming and at times cannot be feasible if the given data set is huge with many missing values.
- ✓ Using a global constant to fill the missing value: This method replaces all missing attribute values with the same constant like a label "Unknown" or $-\infty$. If missing values are replaced with example "Unknown," then the mining program can happen to mistakenly think that they form an interesting concept, because they all have a value in common—that of "Unknown." Though this method is simple and achievable, it is not foolproof.
- ✓ Using the attribute mean or median for all samples that belong to the same class as the given tuple: For instance, if classifying customers based on credit risk, we might replace the missing value with the mean income value for customers who hold the same credit risk category as that of the given tuple. In case the data distribution for a given class is skewed, then the median value is better to be opted.
- ✓ Using the most probable value to fill the missing values: This can be determined with regression, inference-based tools like decision tree induction or Bayesian formalism.

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 result data set has three categories of classes namely Low,Average and High. The result analysis dataset which is shown in Fig-1 is applied to BAYES NET, NAVE BAYES, J48, LOGISTICS, RANDOM FOREST MULTILAYER PERCEPTION, SIMPLE LOGISTIC, SMD, HOEFFIDING TREE and DECISION STUMP to show the accuracy which is shown in Table-2 .

RESULT

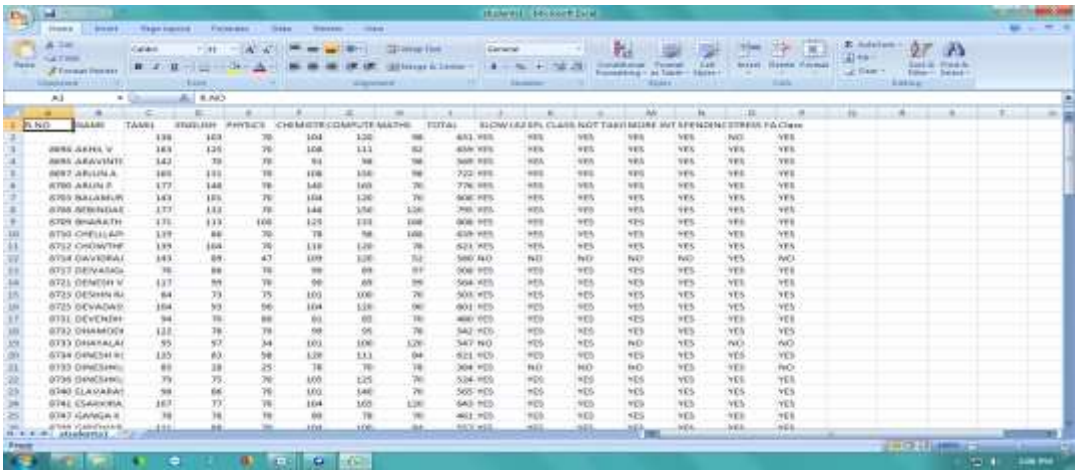

Fig.1. Student Result Analysis Data set in CSV Format

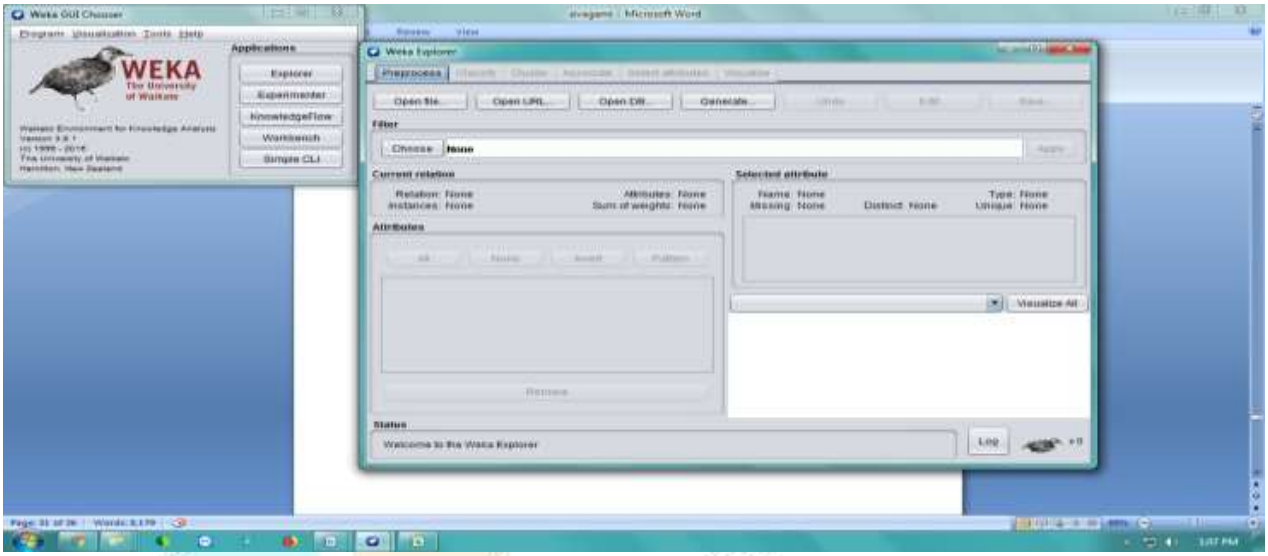

Fig.2 .Weka Explorer Menu

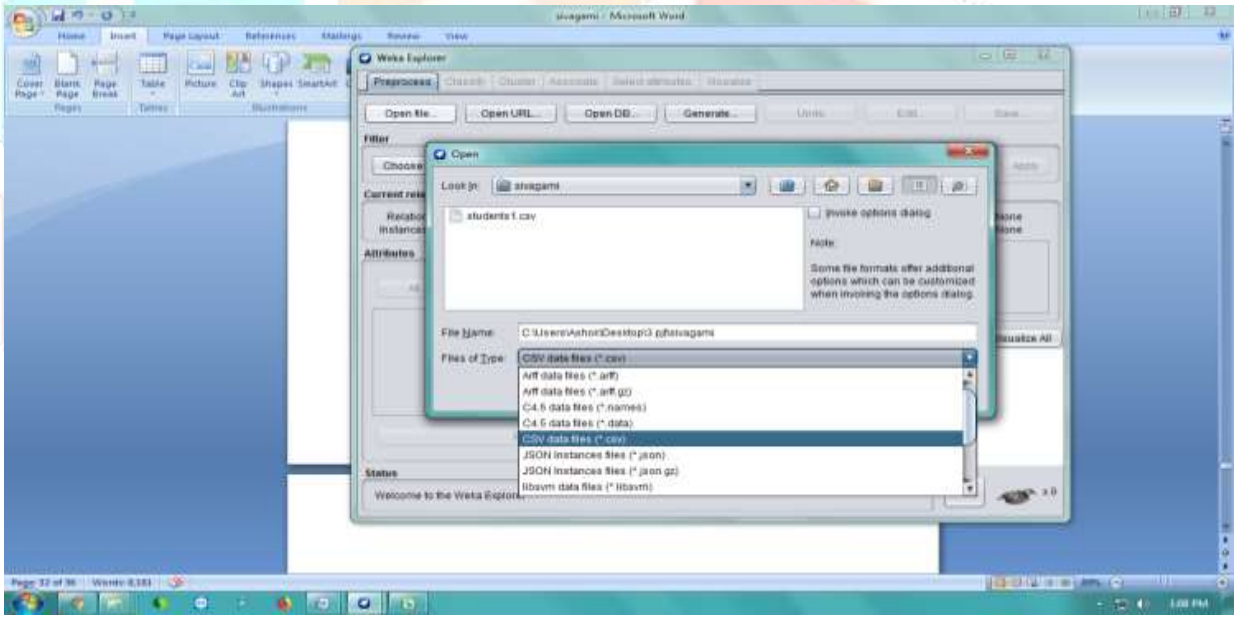

Fig.3.Student Data set Selection

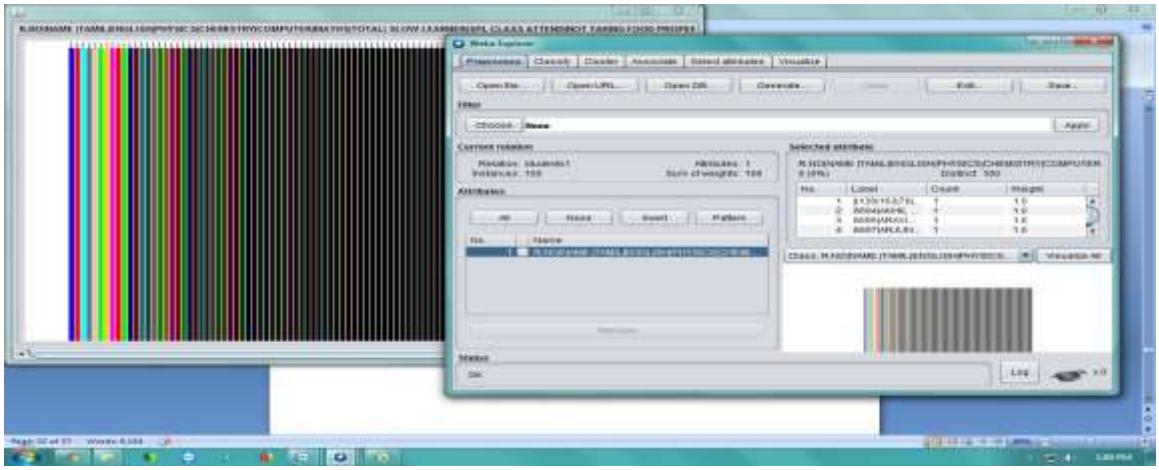

Fig.4.Loaded dataset of students

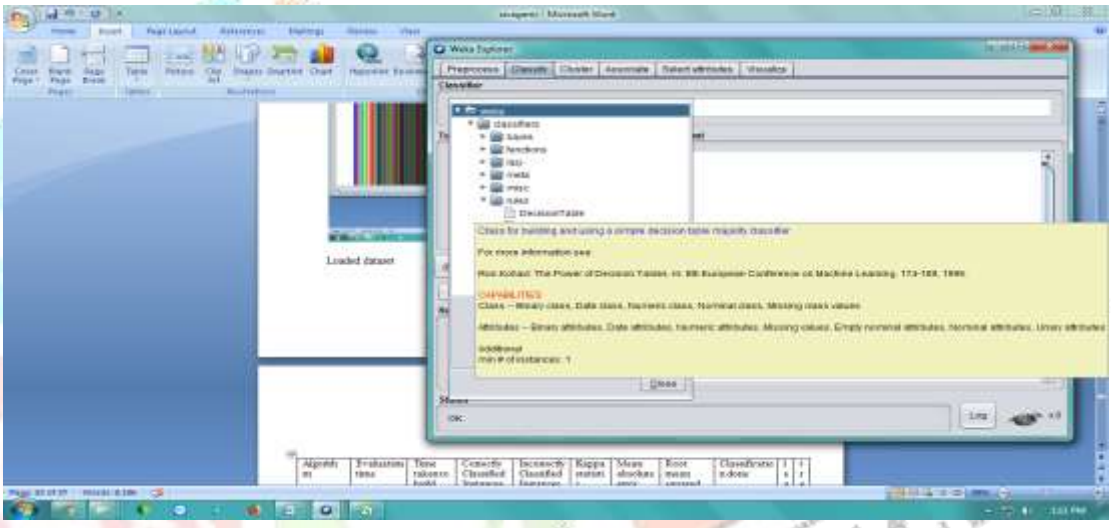

Fig.5.Algorithm selection

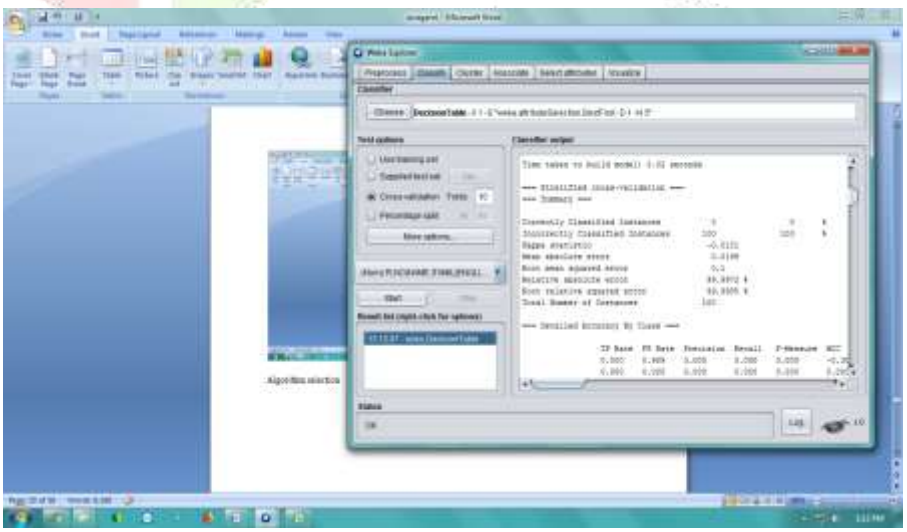

Fig.6.Classification

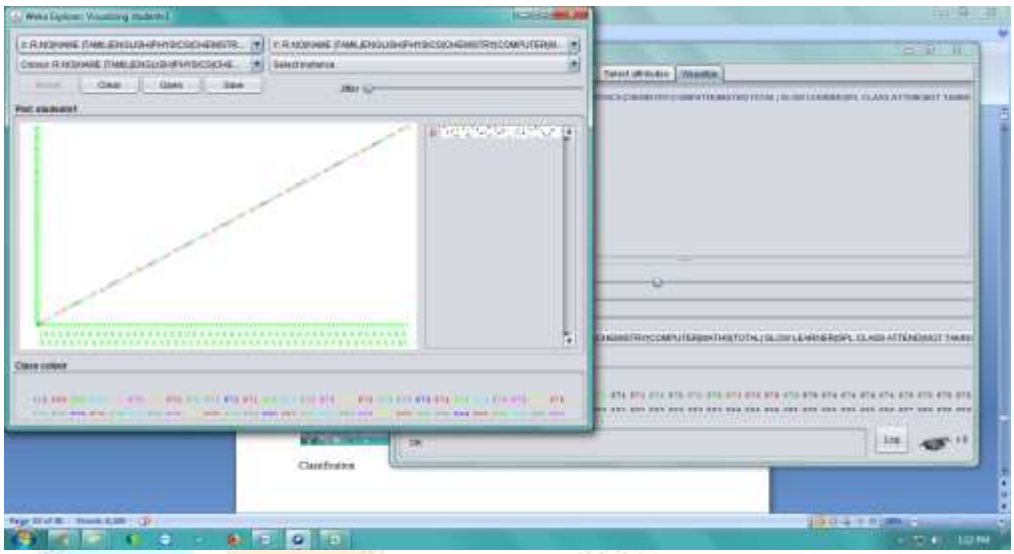

Fig.7.Visualizing the data

Algorithm	Evaluation time	Time taken to build model	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Accuracy
NAVE BAYES	0.01 s	0 s	99%	1 %	0%	0.0198	0.0995	100%
BAYES NET	0.03 s	0 s	99%	1%	0%	0.0198	0.0995	99%
J48	0.01 s	0.01s	97%	3%	0%	0.0198	0.0995	99%
RANDOM FOREST	0.4 s	0.23s	97%	3%	-0.0101%	0.0198	100.0016	98%
LOGISTICS	0.01s	0s	1.7544%	98.2456	0	0.012	0.0776	99%
MULTILAYER PERCEPTION	-	0.01	100%	0%	0	0.012	0.0776	95%
SMD	0.001s	0.02	90%	10%	0	0.012	0.0776	80%
SIMPLE LOGISTIC	0.01s	0s	1.7544	98.2456	0	0.012	0.0776	99%
DECISION STUMP	0.1s	0.1s	100%	0%	0%	0.012	0.0776	94%
HOEFFIDING TREE	0.02s	0.01s	1.7544	98.2456%	0%	0.012	0.0776	99%
LMT	0s	0.29s	99.6988	0.0345%	2%	0.0198	0.0774	91%

Table.1.Comparison of Different Algorithms Using Student Dataset

Conclusion

This paper works on the goal to group the students based on grade order in all their education studies and to improve 'students' academic performance based on the accuracy. The algorithms were applied to the result data set shown in figure-1 signifies that the overall accuracy shown in Table-1 based on the segregation of classes namely Low, Medium and High. The Naïve Bayes shows 100 % accuracy and other algorithms shows 99% accuracy and other algorithms shows below 99% accuracy which signifies that the students performance is based on the subjects namely physics,chemistry,computer science, mathematics, tamil and English and there is a chance that the students may opt for engineering , science or arts courses after completing their high school education.

REFERENCES

1. Mustafa Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," IEEE Access , Volume: 4 ,2016.
2. TriptiMishra,Dr. DharminderKumar,Dr. SangeetaGupta,"Mining Students' Data for Performance Prediction," in fourth International Conference on Advanced Computing & Communication Technologies,2014.
3. Keno C. Piad, MenchitaDumlao, Melvin A. Ballera, Shaneth C. Ambat," Predicting IT Employability Using Data Mining Techniques," in third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 2016.
4. BipinBihariJayasingh,"A Data Mining Approach to Inquiry Based Inductive Learning Practice In Engineering Education," in IEEE 6th International Conference on Advanced Computing,2016.
5. S. M. Merchán,"Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic,"IEEE Latin America Transactions, vol. 14, no. 6, June 2016.
6. KonstantinaChrysafiadi and Maria Virvou," Fuzzy Logic for adaptive instruction in an e-learning environment for computer programming," IEEE Transactions on Fuzzy Systems ,Volume: 23, Issue: 1, Feb. 2015.
7. M. Mayilvaganan,D. Kalpanadevi ," Comparison of Classification Techniques for predicting the performance of Students Academic Environment," in International Conference on Communication and Network Technologies (ICCNT), 2014.
8. Crist´obal Romero," Educational Data Mining: A Review of the State of the Art," IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 40, No. 6, November 2010.
9. PriyankaAnandraoPatil, R. V. Mane," Prediction of Students Performance Using Frequent Pattern Tree," Sixth International Conference on Computational Intelligence and Communication Networks,2014.
[10] Mining Educational Data to Analyze Students" Performance, Brijesh Kumar Baradwaj, Saurabh Pal, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.