ID3 Classifier for Pupils' Status Prediction

K.Nandhini, PhD.
Professor
Department of Computer Applications
Professional Group of Institutions
Coimbatore to Trichy Road
K.N.Puram, Palladam[Tk]-641 662

S. Saranya
M.Phil Scholar
Department of Computer Science
Dr. NGP Arts and Science College
Coimbatore-641 048

ABSTRACT
Predicting the pupil's status is the primary goal. Many studies have been made by a large number of scientists to explore the prediction of their research. One best solution is predicting the results based on the data source by applying some data mining techniques. This research work is to identify the prediction results by means of applying classification technique on the data source being available. There are many approaches in classification technique but this paper implements ID3 (Iterative Dichotomiser 3) Decision Tree concept which provides higher accuracy rates. This model extracts highly useful, reliable patterns from the database to ensure pupil's to achieve a higher academic output.

Keywords
Classification, Decision Tree, ID3 Algorithm, WEKA Tool.

1. INTRODUCTION

Data are collected in various formats like images, audios, videos, records, text files, etc. are stored in the database. These data can be efficiently extracted and mined from the repository by means of Knowledge Discovery Process often called Data Mining, so that it can be used to provide a proven solution to the problem defined. In simple words KDD can be defined as the process of extracting the useful and valid patterns from the data warehouse for resolving a problem. Data mining is being used in many areas like banking, insurance, hospital, education and in new fields of Statistics, Databases, Machine Learning, Pattern Recognition, Artificial Intelligence (AI) and Computation capabilities etc. Nowadays data mining has extended its research in the field of education.

Educational Data Mining (EDM) uses various techniques like Classification, Decision Trees, Rule Induction, Nearest Neighbors, Neural Networks, Clustering, Genetic Algorithms, Exploratory Factor Analysis and Stepwise Regression. Each technique provides different sort of results, this work computes the result by means of ID3 algorithm of decision tree. Information like Personal details, Semester percentage, Extracurricular activities, Seminar participation, Paper presentation, Certification courses etc. based on these discipline and behavior were collected from the pupil's management system, to predict the status. This paper investigates the accuracy of Decision tree techniques for predicting pupil's performance by using the Data Mining Machine Learning Tool WEKA. Performance profiling is dependent upon the motivation, attitude and marks and by the continued real-time monitoring of pupil's performance using a simple rapid response system that predicts correctly which pupil may need some extra attention in the course of their education. The model developed to achieve a measurable pupil's progress monitoring process that gives results quickly and meets a larger educational goal benefiting stakeholders in the educational system and the wider community.

1.1 WEKA TOOL

WEKA is an open source java code created by researchers at the University of Waikato in New Zealand. WEKA is a library of Java routines for various machine learning tasks. WEKA can also be used as a stand-alone learner. It provides many different machine learning algorithms, including Decision Tree (J4.8, an extension of C4.5), MLP- Multiple Layer Perceptron (a type of neural net), Naïve Bayes, Rule Induction Algorithms such as JRip, Support Vector Machine (SVM) and more.

Data mining is solely the domain of big companies and expensive software. In fact, there is a piece of software that does almost all the same things as these expensive pieces of software, called WEKA. It uses the GNU General Public License (GPL). The software is written in the Java language and contains a GUI for interacting with data files and producing visual results. It also has a general API, so you can embed WEKA, like any other library, in our own applications to such things as automated server side data mining tasks.

WEKA is preferred method for loading data is in the Attribute Relation File Format (ARFF). First define the type of data being loaded, supply the data itself. In the file, define what each column contains. In the case of the regression model, it's limited to a NUMERIC or a DATE column. Finally, supply each row of data in a Comma Delimited Format.

The main objective of this paper is to use data mining methodologies to study pupil's performance in the courses. Data mining widely used to study pupil's performance. In this paper Decision Tree Method is used for data classification. Information like personal details, percentage, extracurricular activities were collected from the department of computer science to evaluate the pupil's status. This research paper investigates the prediction of pupil's status by applying ID3 classifier technique.

2. LITERATURE SURVEY

Data mining has evolved its research very well in the field of education in a massive amount. This tremendous growth is mainly because it contributes much to the educational systems to analyze and improve the growth of pupil's as well as the pattern of education.

Han and Kamber [1] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.

Galit [2] gave a case study that use pupil's data to analyze their learning behavior to predict the results and to warn pupil's at risk before their final exams.

Al-Radaideh, et al [3] applied a decision tree model to predict the final grade of pupil's who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Khan [4] conducted a performance study on 400 pupil's comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Hijazi and Naqvi [5] conducted study on the pupils' performance by selecting a sample of 300 pupils' (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Pupils' attitude towards attendance in class, hours spent in study on daily basis after college, pupils' family income, pupils' mother's age and mother's education are significantly related with pupils' performance" was framed. By means of simple linear regression analysis, it was found that the factors like mother's education and pupils' family income were highly correlated with the pupils' academic performance.

Cortez and Silva [6] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school pupils' from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 pupils', who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (*pass/fail*) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.
Pandey and Pal [7] conducted study on the pupils' performance based by selecting 600 pupils' from different colleges of Dr. R. M. L. Awadh University, Faizabad, India.

By means of Bayes Classification on category, language and background qualification, it was found that whether new comer pupils' will performer or not.

Bray [8], in his study on private tutoring and its implications, observed that the percentage of pupils' receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Srilanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

4. DATA SOURCE

The pupil's status is determined mainly by their semester percentage, extracurricular activities, seminar presentation, certification course and paper presentation. Each pupil's has to get minimum marks to pass a semester in internal as well as semester examination. The data set used in this study was obtained from Dr. N.G.P arts and science college, Coimbatore (Tamil Nadu) on the sampling method of computer science department of course M.Sc. (Master of Computer Science) from session 2011 to 2013. Initially size of the data is 60. By

3. ID3 (Iterative Dichotomiser 3)

ID3 developed at the University of Sydney. Ross Quinlan first presented ID3 in 1975 in a book, Machine Learning, vol. 1, no. 1. ID3 is based on the Concept Learning System (CLS) Algorithm [9].

3.1 Splitting Criteria

A fundamental part of any algorithm that constructs a decision tree from a dataset is the method in which it selects attributes at each node of the tree.

3.2 Entropy

A measure used from Information Theory in the ID3 algorithm and many others used in decision tree construction is that of Entropy. Informally, the entropy of a dataset can be considered to be how disordered it is. It has been shown that entropy is related to information, in the sense that the higher the entropy, or uncertainty, of some data, then the more information is required in order to completely describe that data. In building a decision tree, we aim to decrease the entropy of the dataset until we reach leaf nodes at which point the subset that we are left with is pure, or has zero entropy and represents instances all of one class (all instances have the same value for the target attribute). We measure the entropy of a dataset, S, with respect to one attribute, in this case the target attribute, with the following calculation:

Entropy measures the amount of information in an attribute. Given a collection S of c outcomes Entropy(S) = S -p(I) log2 p(I) where p(I) is the proportion of belonging to class I. S is over c. Log2 is log base 2.Note that S is not an attribute but the entire sample set.

If S is a collection of 14sample records with 9 YES and 5 NO sample then Entropy(S) = - (9/14) Log2 (9/14) - (5/14) Log2 (5/14) = 0.940. Notice entropy is 0 if all members of S belong to the same class (the data is perfectly classified). The range of entropy is 0 ("perfectly classified") to 1 ("totally random").

3.3 Gain

Gain is computed to estimate the gain produced by a split over an attribute, Gain (S, A) is information gain of example set S on attribute A is defined as
Gain(S, A) = Entropy(S) - S (($|S_v|$ / $|S|$) * Entropy (S_v))

Where:
S is each value v of all possible values of attribute A

S_v = subset of S for which attribute A has value v
$|S_v|$ = number of elements in S_v
$|S|$ = number of elements in S

Gain quantifies the entropy improvement by splitting over an attribute: higher is better.

identifying these pupils' known, we are able to decide which type of pupils' are more successful than others and provide academic help for those who are less likely to be successful.

4.1 Data Preparation

During data collection, the relevant data is gathered and the quality of data must be verified. Usually, the assembled data contains of missing or incomplete attribute, noisy (containing errors, or outlier values that deviate from expected), and inconsistent of data are common. Therefore, the collected data must be cleaned and transformed before it can be utilized in data mining system since data mining should process cleaned data in order to come out with better and or quality results. Data cleaning involves several of processes such as filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Then, the cleaned data's are transformed into a form of table that is suitable for data mining model. The cleaned data will be divided into two; training or learning data (60%) and the rest is for validating the data. These training data is applied to develop the model while the validated data is used to verify the chosen model.

5. RESULTS

The data set of 60 pupils' used in this study was obtained from Dr.NGP Arts and Science College, Coimbatore (Tamil Nadu) Computer Science Department of course MSc (Master of Computer Science) from session 2011-2013.

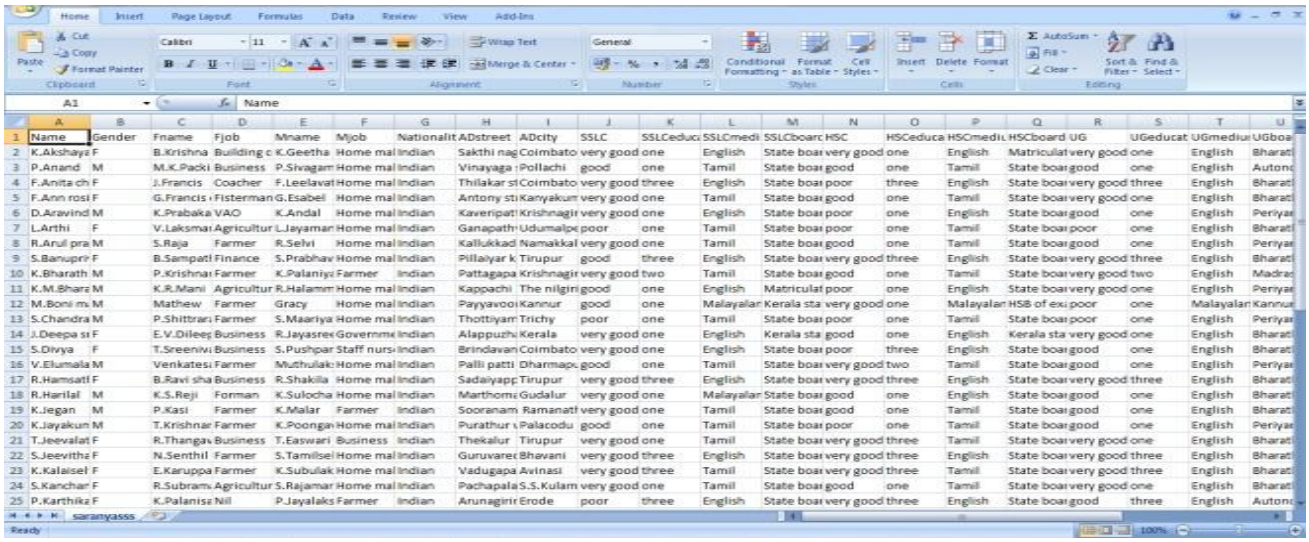

Fig 1: Dataset

The above fig 1 contains 60 pupils' personal details of their SSLC, HSC, UG degree percentage and family details like their parents education qualification to determine whether they are eligible to do higher studies or not and also able to determine the pupils' status. In this paper using WEKA tool to find out the eligible pupils' by using ID3 classifier.

The following step clearly specifies the ID3 classifier performance in WEKA tool.

Step-1: first we have to import the data's into the WEKA explorer and then preprocess the data. The following screen identifies the preprocessing step.

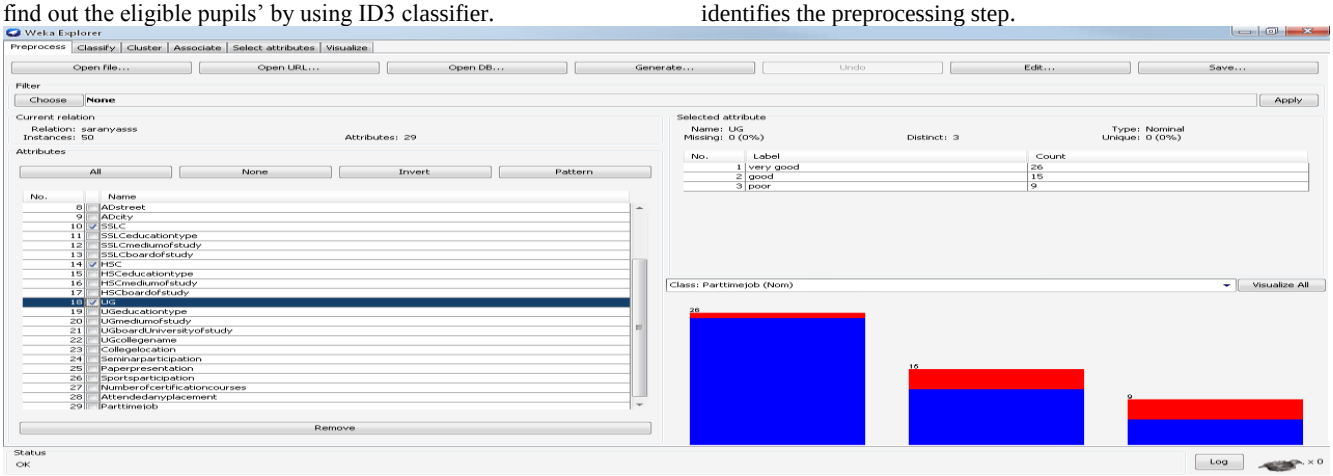

Fig 2: Preprocessing

Step-2: next classify the data by entering the cross validation value as 10. The screen will be like this,

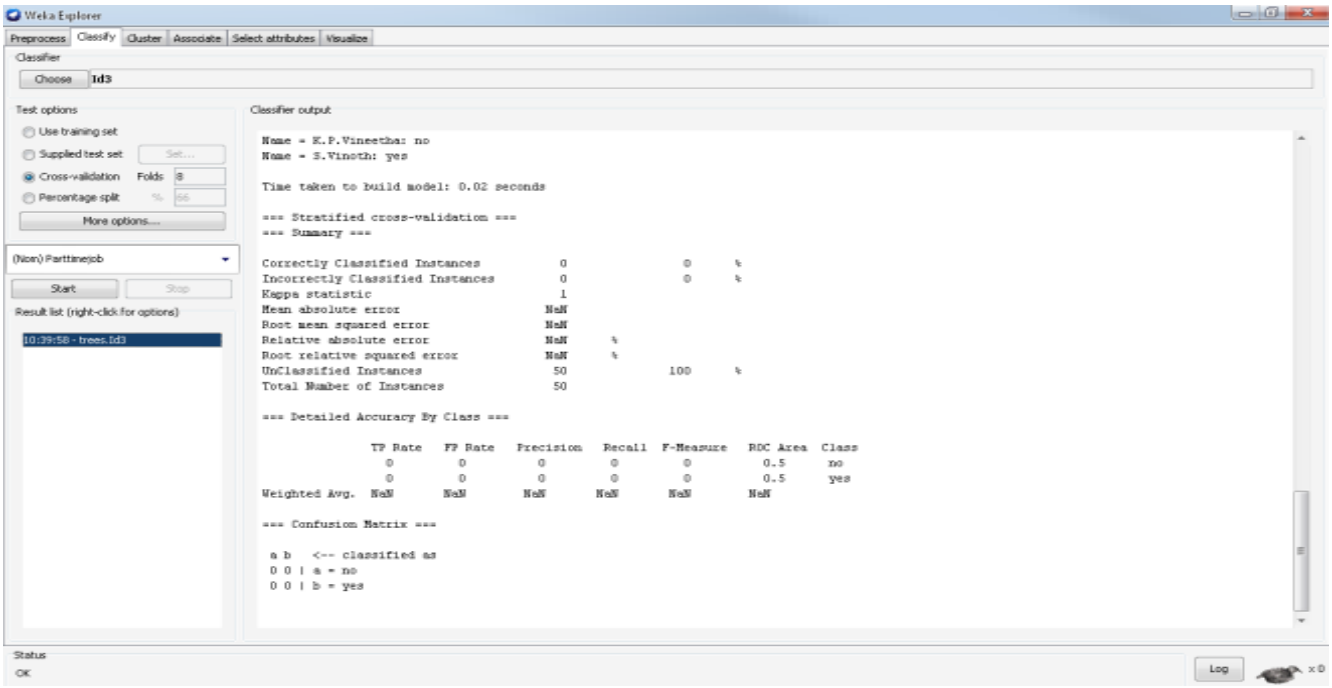

Fig 3: Classification

Step-3: third step selects the attributes which gives the first priority to derive the tree whether as a left node or right node.

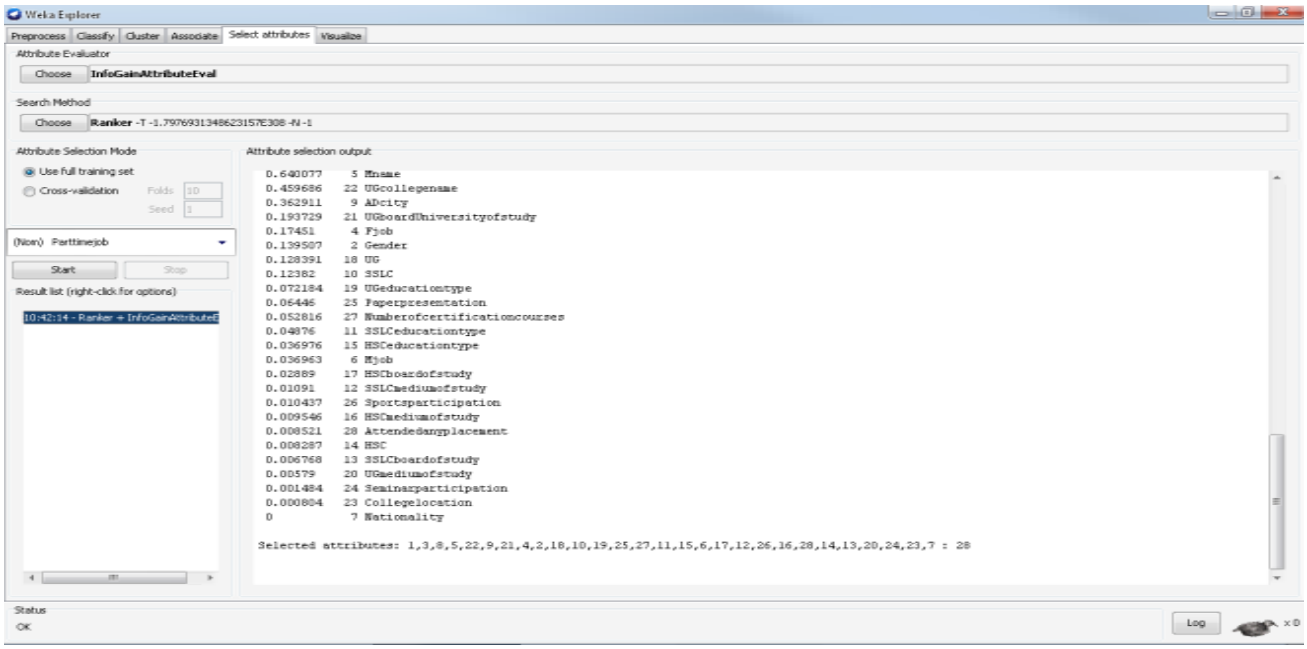

Fig 4: Select Attribute

Step-4: The following diagram specifies the overall data set visualization.

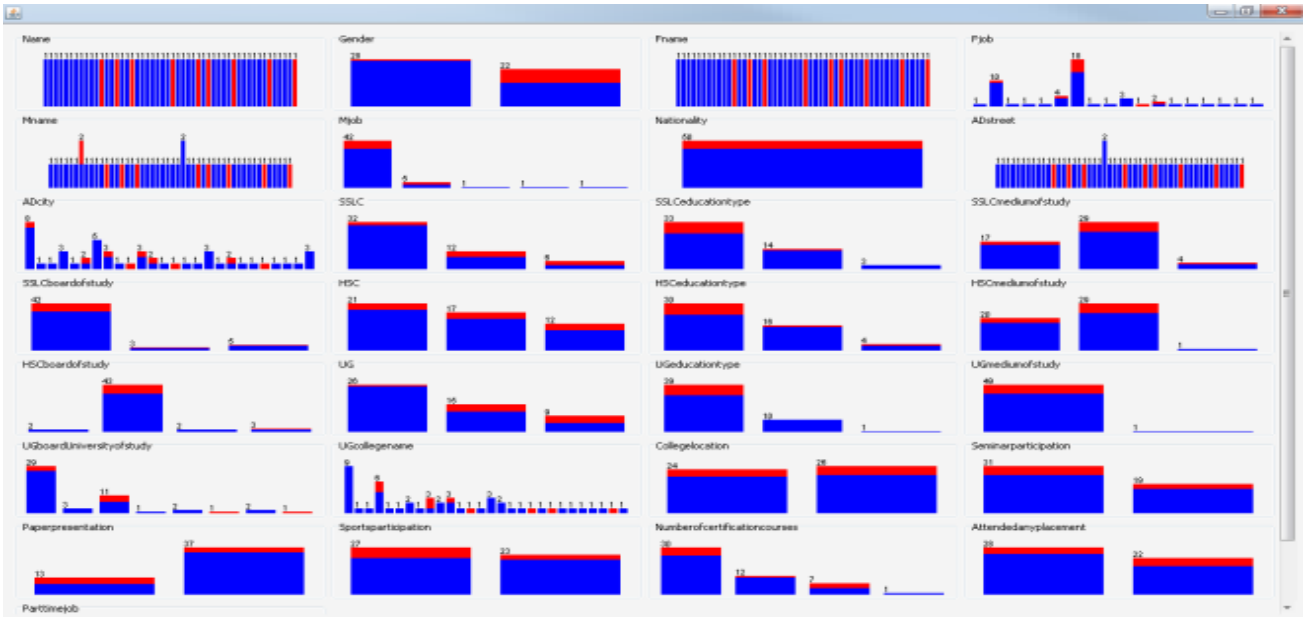

Fig 5: Visualization

6. CONCLUSION AND FUTURE ENHANCEMENT

This paper analogous classification techniques which is ID3 classifier by using the data mining tool WEKA. The ID3 classifier used to predicting the right pupil's who are eligible to do their PG courses. The efficiency of the ID3 algorithm is highly accurate than the other mentioned algorithms. This paper analyses which pupil performing good or bad in their upcoming courses. The above steps denote the ID3 performance using WEKA. In future the ID3 algorithm combined with genetic algorithm to predicting the pupil's status.

7. REFERENCES

[1] Andrew Colin, "Building Decision Trees with the ID3 Algorithm", by: Dr. Dobbs Journal, June 1996.

[2] Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.

[3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[4] M. Bray, The Shadow Education System: Private Tutoring and Its Implications for Planners, (2nd ed.), UNESCO, PARIS, FRANCE, 2007.

[5] P. Cortez, and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J.Teixeira (Eds.), 2008, pp.5-12.

[6] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[7] S. T. Hijazi, and R. S. M. M. Naqvi, "Factors Affecting Student's Performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[8] U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[9] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp.84-87.