International Journal of Engineering and Techniques - Volume 3 Issue 5, Sep - Oct 2017

RESEARCH ARTICLE OPEN ACCESS

Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School

Paul Joseph M. Estrera , Pamela E. Natan, Babe Grrece T. Rivera, Faith B. Colarte
Department of Information Technology
College of Information Technology and Computing
University of Science and Technology of Southern Philippines
C.M. Recto Avenue, Lapasan, Cagayan de Oro City, 9000 Philippine

Abstract:

Predicting students' academic performance is mostly useful to help the educators and learners improve their teaching and learning process. In this study, the researchers develop a system that merges the work of a dynamic web-based grade book and predictive analytics for the students' performance. We have performed series of test to evaluate the importance of the attributes and the results shows that the student's GPA, gender, study behavior, interest in studies and the engaged time in studying of the USTP Senior High School students had a great impact in the prediction. The researchers also found out that using the Decision Tree Algorithm is efficient for the prediction. It can be concluded that this paper will help the students and teachers to monitor the students' performance in a systematic way and taking appropriate action to improve it. The researchers recommend that the students' entrance exam results and extra-curricular activities would be included in predicting the academic performance of the students.

Keyword **Gradebook, Attributes, Predictive Analytics, Decision Tree, Student Performance**

I. INTRODUCTION

A. *Background and Rationale*

Educational institutions aim to provide quality education and analyze the performance of students and help them improve. The varying factors in current education has led to the pursuit of effective and efficient monitoring of student performance, thus, the ability to predict students' performance serves a vital role in providing information that is geared to help students, teachers, administrators, and policy makers take decisions, as in [3].

Student's performance is an essential part in higher learning institutions. This is because one of the criteria for a high quality university is based on its excellent record of academic achievements (M. of Education Malaysia, 2015), furthermore, prediction of student performance helps in providing students with the necessary assistance in the learning process [1]. In order to improve students' achievement and success more effectively in an efficient way by means of predicting students' performance, can be attained using educational data mining techniques which could bring the benefits and impacts to students, educators and academic institutions.

In recent years, there has been an increased interest in using data mining for educational purposes. There are many techniques in data mining that can be applied to educational data, such as K-Nearest Neighbor, Decision Trees, and Naïve Bayes to name a few. The process stated which is the educational data mining is being used for studying the data available in the educational field and bring out the hidden knowledge from it.

Various studies have been conducted which addresses the concern of prediction of students' performance. Several approach, techniques, and factors with regards to data mining have been conducted in order to come up with the prediction. However, based on our research, there are no studies conducted to predict the student's probability of becoming an honor student.

The researchers would like to propose a system that would merge the work of a dynamic web-based grade book and predictive analytics of the students' performance using the RapidMiner Studio, a tool in data mining that would create the Decision Tree algorithm. For this purpose, we have analyzed the data of the Senior High School Students in University of Science and Technology of Southern Philippines. This data was obtained from the information provided by the admitted students using a survey questionnaire that was conducted on them. It includes their gender, age, grade and section, interest in studies, rating of their study behaviour, the time engage in studying, their family support, and their GPA. the following attributes will then be used for the prediction.

The gradebook serves a vital role on the system for it is one of the components that is needed for the computation of the student's GPA, the reason why most researchers are using GPA is its concrete value for future educational and career mobility and an indication of comprehensive academic potential (U. bin Mat *et al*., 2013), along with the other factors which includes the student's demographics such as gender was also used for predicting student performance. Student's psychometric factor is another factor which is identified as student interest, study behavior, engage time, and family support, extra-curricular activities, community involvement, parent educational background and family size.

The proposed system would like to address the concern on how a certain student can monitor its academic performance, to determine he's current class standing so that they would be aware on how well they are doing on their class regarding their academic performance, grades, where he excels and his weaknesses, and specifically to analyze whether he is capable of becoming an honor student which serves basis from the said factors.

B. Review of Related Literature

1) *Data Mining: A prediction for performance improvement using classification:* Reference [2] described the process of knowledge discovery from databases using a practical example of a current actual problem. They developed two models based on decision tree which were successfully used to predict student success based on GPA criterion and time student needs to finish the undergraduate program (time-to-degree) criterion.

2) *A Review on Predicting Student's Performance using Data Mining Techniques:* It is a systematical literature review on predicting student performance by using data mining techniques that is proposed to improve student's achievements. It provides an overview on the data mining techniques that have been used to predict student's performance. It focuses on how the prediction algorithm can be used to identify the most important attributes in a student's data.

2.1 Important Factors on Predicting Student's Performance

There are two main factors in predicting student's performances, which are attributes and prediction methods. First step will be focused on the important attributes used in predicting student performance and second step will be focused on the prediction methods used in predicting student's performance.

The systematical literature review is used to identify the important attributes in predicting student's performance. The attributes that have been frequently used is cumulative grade point average (CGPA). Next, the most often attribute being used is student's demographic and psychometric factors. The students demographic include gender and family support while psychometric factors are identified as the interest of a student in their studies, their study behaviour and their engaged time in studying. The extra-curricular activities are also used in predicting the student performance.

2.2 The prediction methods used for student performance

In educational data mining method, predictive modeling is usually used in predicting student performance. In order to build the predictive modeling, there are several tasks used, which are classification, regression and categorization. The most popular task to predict student's performance is classification. There are several algorithms under classification task that have been applied to predict student's performance. Among the algorithms used are Decision Tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine.

2.2.1 Decision Tree

Decision Tree is one of a popular technique for prediction. Most of researchers have used this technique because of its simplicity and comprehensibility to uncover small or large data structure and predict the value.

2.2.2 Naive Bayes

Naive Bayes algorithm is also an option for researchers to make a prediction. Some research showed that Naive Bayes has used all of attributes contained in the data.Then, it analyzed each one of them to show the importance and independency of each attributes.

2.2.3 K-Nearest Neighbor

K-Nearest Neighbor gave the best performance with the good accuracy. According to Bigdoli et al. (2003), K-Nearest Neighbor method had taken less time to identify the student's performance as a slow learner, average learner, good learner and excellent learner. K-Nearest Neighbor gives a good accuracy in estimating the detailed pattern for learner's progression in tertiary education.

2.3 Discussion

The results analysis of the recent works in predicting student's performance is shown in the figure. This meta-analysis is based on the highest accuracy of prediction methods and also the main important factors that may influence the student's performance. This shows the prediction accuracy that uses classification method grouped by algorithm for predicting student's performance.

3) *A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction:* Data mining helps to extract the relevant information from the large and complex databases. Data mining techniques are useful for data analysis and predictions. Classification is an unsupervised learning technique that helps to classify predefined class labels. There are various classification techniques such as Decision tree algorithm, Bayesian network, Neural network and Genetic algorithm etc. These technique can be used to build the classification model. This classification model helps to predict the future trend based on previous pattern. This paper propose a classification model particularly decision tree algorithm to predict the future grades of the students in their final examinations. A number of factors may affect the performance of students. Here some significant factors have been considered while constructing the decision tree for classifying students according to their attributes (grades). In this paper four different decision tree algorithms J48, NBtree, Reptree and Simple cart were compared and J48 decision tree algorithm is found to be the best suitable algorithm for model construction.

A classification model has been proposed in this study for predicting student's grades particularly for engineering under graduate students. Four decision tree algorithms were compared and J48 decision tree algorithm was selected for model construction, where J48 is a java version of C 4.5.The model obtained accuracy of 80.15% and 82.58% in 10 fold cross validation method and percentage method respectively. It indicates that model is good for forecasting the grades of students. This model helps to the management to identify weak students and can take from failure.

4) *A Framework for Students Academic Performance Monitoring and Evaluation System (SAPMES) For Higher Education Institutions:* SAPMES is a performance evaluation and monitoring system for HEI that provides the students 24/7 online access of their academic record and can monitor the progress of term grades and final grades via internet. It also provides the parents a way to track their child's academic performance. Since SAPMES is available online, making updates of the class record is easy and all information in the system can be exported to spreadsheet format so faculty can have a softcopy of the grade book and a hardcopy when printed. It also supports transparency of student's academic records. Progress of classroom activities that the faculty undertakes can also be monitored by the administrator [7].

5) *Data Mining Applications: A comparative Study for Predicting Student's performance:* Reference [9] concluded that Decision trees are so popular because they produce classification rules that are easy to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Student data to predict the student's performance in the end semester examination. The experimental results show that Classification and Regression Tree (CART) CART is the best algorithm for classification of data. From the study conducted by (Burnette, 2013), by using a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. Two algorithms CHAID and CART were applied on student enrolment data of information system students of open polytechnic of New Zealand

to get two decision trees classifying successful and unsuccessful students. The result obtained showed that the accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

6) *ID3 (Iterative Dichotomiser 3) Decision Tree algorithm:* ID3 (Iterative Dichotomiser 3) Decision Tree algorithm is one of the classification of the Decision Tree. The Iterative Dichotomiser 3 Decision Tree algorithm is invented by Ross Quinlan used to generate a decision tree from the dataset. The decision tree technique involves constructing a tree to model the classification process. Once a tree is built, it is applied to each tuple in the database and the results in classification for that tuple. The Iterative Dichotomiser 3 Decision Tree algorithm is a classification algorithm based on Information Entropy.

The following issues are faced by most decision tree algorithms [4]:

- Choosing splitting attributes
- Ordering of splitting attributes
- Number of splits to take
- Balance of tree structure and pruning

II. METHODOLOGY

This chapter presents the process in developing the system namely, design analysis, database design, system design, and development of the system and evaluation of the system.

A. *Analysis*

Through thorough study on student performance prediction on previous researches, the researchers designed a dynamic grade book system specifically designed for Senior High Students of USTP that has the capability to predict its students' academic excellence using a predictive algorithm. We created a simple and user-friendly user interface that will help the teachers to compute grades automatically and see their students' academic performance.

B. *Important Factors on Predicting Student's Performance*

There are two main factors in predicting the student's performance; these are the attributes and its prediction method. The first step will be focused on the important attributes used in predicting student performance and the second step will be focused on the prediction algorithm used.

1) *Attributes:* An intensive review was conducted to identify the important attributes in predicting student's academic performance. Five attributes were selected for the mining process based on the literature review done on previous work as shown on Table 1.

TABLE I
RESULT ACCURACY USING DECISION TREE METHOD (A REVIEW ON PREDICTING STUDENT'S PERFORMANCE USING DATA MINING TECHNIQUES, 2013 [11])

Method	Attribute	Accuracy	Authors
Decision Tree	GPA	91%	Jishan et al. (2015)
	Psychometric factors	65%	Gray et al. (2014)
	Extra-curricular Activities	73%	Mishra et al. (2014)
	Student Demographic	65%	Rameshet al. (2013)

The first attribute was the General Average of the student. Brown (1966) reports that high school GPA adequately predicts success in future courses.

The next attribute is the student demographic. Student demographic includes gender. The reason why most of the researchers used student's gender is because they have different styles of female and male students in their learning process. Study done by Meit et al. (2007) found that most of female students have various positive learning styles and behaviors compared to male students. Female students are more discipline and dutiful in their studies, self-directed, always preserved and focused. Thus, it is proven that gender is one of important attributes influencing student's performance.

There are also several researchers in another study who have used psychometric factor to predict student's performance. A psychometric factor is identified as student interest in studies, its rating in their study behavior, engage time in studying, family support, community involvement, parent educational background and family size. It helps the lecturer to evaluate students achievement based on their personal interest and behavior.

Table 2 summarizes the actual parameters and its corresponding categorical value that is used in predicting student academic performance.

TABLE 2
PARAMETERS USED FOR STUDENT PERFORMANCE PREDICTION

Parameters	Description	Possible Values
GPA	Student's Grade Point Average	Excellent (98-100), Good (95-97), Fair (90-94), Poor (90 below)
Gender	Student's Gender	Male, Female
Student's interest	Does the Student has interest in studies	Yes, No
study behaviour	Student's study behavior	Low, Medium, High
family support	Does the family support the student	Yes, No
engage time	Engaged time in studying	High, Low
Community involvement	Student's Community Involvement	Yes, No
Parent education	Student's parent education attainment	Elementary Graduate, High School Graduate, College Graduate
Family size	student's family size	Big, Small

2) *Prediction Algorithm:* To have a better insight on determining which predictive algorithm is best to use, the performance of the ID3 Decision Tree, Naïve Bayes and K-Nearest Neighbour algorithms are examined by evaluating the accuracy of the results on the datasets we have gathered. The Decision Tree shows the highest accuracy value of 91% compared to the other techniques. The Naïve Bayes was next with an accuracy of 85.97%, and lastly the K-Nearest Neighbor with an accuracy of 58%. The results are summarized in Fig. 1. The result was evaluated using a 10-fold cross validation. In 10-fold cross-validation, the data set is randomly partitioned into 10 equal size subsets. Of the 10 subsets, a single subset is retained as the test data for testing the model, and the remaining 9 subsets are used as training data. The cross-validation process is then iterated 10 times (the folds), with each of the 10 subsets used exactly once as the validation data. The 10 results from the folds arethen computed their average to produce a single estimation. A diagram is provided in Fig. 2.

Fig. 1 Prediction Accuracy Grouped by Algorithms

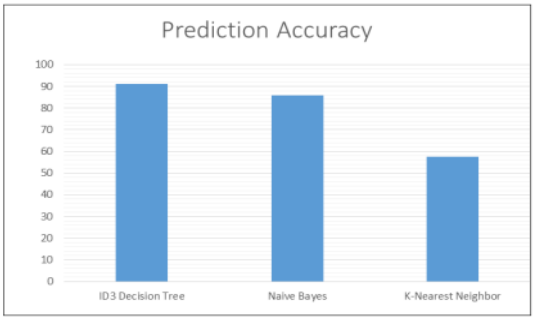
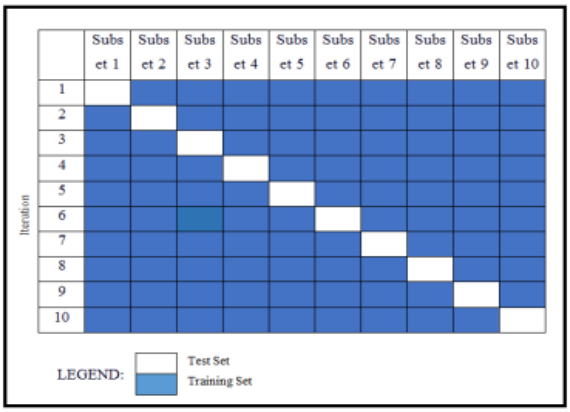


Fig. 2 Ten Fold Cross Validation



With the highest accuracy value of 91%, the Decision Tree Algorithm will be used in predicting the students' academic performance. Specifically, the ID3 (Iterative Dichotomiser 3) Decision Tree algorithm will be used for the prediction.

C. *Framework for Predicting Student Academic Performance*

The framework in processing a prediction model by using some data mining techniques is shown in Fig. 3. It illustrates the three main stages involved in this study.

1) *Data Collection:* The 150 students' data for the prediction model were collected through a questionnaire survey conducted at University of Science and Technology of Southern Philippines academic year 2016-2017, among the Grade 11 students and was filled into a MySQL database. These data includes their gender, age, grade and section, interest in studies, rating of their study behaviour, the time engage in studying, their family support, and their GPA.
2) *Data Transformation:*The data transformation stage was performed to improve the quality of input data to produce better results. Once the details of all the students are collected, it was then segmented further,

considering various feasible attributes which would have a higher impact on the performance of a student and irrelevant, missing and incomplete attributes had been removed. The attributes that had been retained are GPA, gender, student's interest, study behavior, family support, engage time, Failing grade and Extracurricular. Finally, the "result" attribute was added and it held the predicted result, which can be either "With Honor", "With High Honor", "With Highest Honor" or "not an honor student".

3) *Algorithm Implementation:* After the data was cleaned, it was then processed in the RapidMiner. The cleaned data was cross validated. It was randomly partitioned into training test and test set and the Decision Tree Method was applied. The results were evaluated and the Decision Tree Model was then generated.

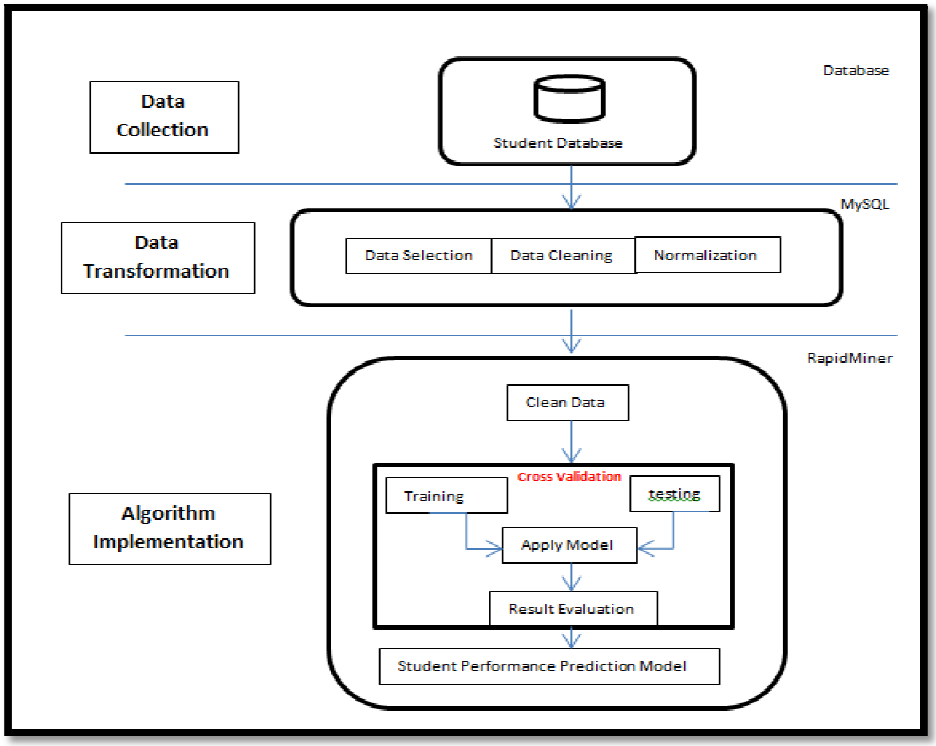


Fig. 3 Processing Model

D. Software Programming Tools

1) *HTML and CSS:* HyperText Markup Language (HTML) is a markup language for creating web pages or other information to display in a web browser. HTML allows images and objects to be included and that can be used to create interactive forms. From this, structured documents are created by using structural semantics for text such as headings, links, lists, paragraphs, quotes etc. CSS (Cascading Style Sheets) is designed to enable the separation between document content (in HTML or similar markup languages) and document presentation. This technique is used to improve content accessibility also to provide more flexibility and control in the specification of content and presentation characteristics. This enables multiple pages to share formatting and reduce redundancies. We have use the HTML and CSS in designing our Grade book system.

2) *PHP and the CodeIgniter Framework:* PHP (recursive acronym for PHP: Hypertext Preprocessor) is a widely-used open source general purpose server side scripting language that is especially suited for web development and can be embedded into HTML. CodeIgniter is a well-known open source web application framework used for building dynamic web applications in PHP [6]. Its goal is to enable developers to develop projects quickly by providing a rich set of libraries and functionalities for commonly used tasks with a simple interface and logical structure for accessing these libraries. CodeIgniter is loosely based on the Model-View-Controller (MVC) pattern and we have used it to build the front end of our implementation. We have use the PHP language

and codeigniter framework in developing our web based gradebook system.

3) *MySQL:* MySQL is the most popular open source RDBMS which is supported, distributed and developed by Oracle [8]. In the implementation of our web application, we have used it to store user information and students' data.

4) *RapidMiner Studio and Server:* RapidMiner is an open source data mining tool that provides data mining and machine learning procedures including data loading and transformation, data preprocessing and visualization, modelling, evaluation, and deployment. It is written in the Java programming language and makes use of learning schemes and attribute evaluators from the WEKA machine learning environment and statistical modelling schemes for the R-Project. We will use RapidMiner to create model for the prediction. Several Operators in RapidMiner Studio was used in creating a training vector for the prediction such as the Read Database operator, Select Attributes operator, Set Role operator, Multiply Operator, Cross Validation operator, Filter Examples operator, Apply Model operator, ID3 operator, Performance Classification operator, Rename Operator, Execute SQL operator and Update Database operator. The performance Classification Operator will give the accuracy of the prediction algorithm.

III. RESULTS AND DISCUSSION

To get a better insight into the importance of the input variables, it is customary to analyze the impact of input variables during students' prediction success. The impact of certain input variable of the model on the output variable has been analyzed. Tests were conducted using three tests for the assessment of input variables: Chi-square statistics test, Information Gain test and Information Gain Ratio test. The results obtained with these values are shown in Table 3.

TABLE 3

THE RESULT OF ALL TESTS AND ITS AVERAGE RANK

The aim of this analysis is to determine the importance of each attribute individually, and the higher the weight of an attribute, the more relevant it is considered. Table 4 shows that the attribute GPA impacts the output the most, and that it showed the best performances in all of the four tests. Then these attributes follow: Behavior, Time, Support, Interest, Gender, community involvement, parent education and family size. Since the Community involvement, parent education and family size showed an average of 0 in all the test performed, thus the three attributes are eliminated in predicting the students' performance.

Rank	Attribute	Information Gain	Information Gain Ratio	Chi squared statistics	Average
1	gpa	1.25	0.87	423.81	141.98
2	behaviour	0.07	0.05	10.29	3.47
3	time	0.06	0.06	10.88	3.67
4	support	0.02	0.03	3.35	1.13
5	interest	0.01	0.01	1.09	0.37
6	Comminvolve	0.00	0.00	0.01	0.00
7	parenteduc	0.00	0.00	0.02	0.00
8	familysize	0.00	0.00	0.01	0.00

We have also carried out some experiments in order to evaluate the performance and usefulness of the Prediction Algorithm in predicting the students' academic excellence. The results of the experiments are summarized in Table 4. The performance of the prediction algorithms are evaluated based on its prediction accuracy.

TABLE 4

PREDICTIVE PERFORMANCE OF THE PREDICTION ALGORITHM

Evaluation Criteria	Decision Tree Algorithm	Naïve Bayes	K-Nearest Neighbor
Prediction Accuracy	90.67%	85.97%	58%

The Decision Tree method shows a (90.67%) of performance accuracy. The Naïve Bayes method shows a (85.97%) of performance accuracy while K-Nearest Neighbor shows a (58%) of performance accuracy.

Figure 4 shows the generated decision tree after running the RapidMiner. This shows that the GPA is the root node for the tree. If the GPA is excellent then it will be followed by its interest if it is a yes then he will be a with highest honor. And if it's a no, then he will be a with high honor. Now if its GPA is fair then the next node is the time spent in studying. If the GPA is good, the next node is its rate in interest of studying and if the GPA is poor, then he is not an honor student.

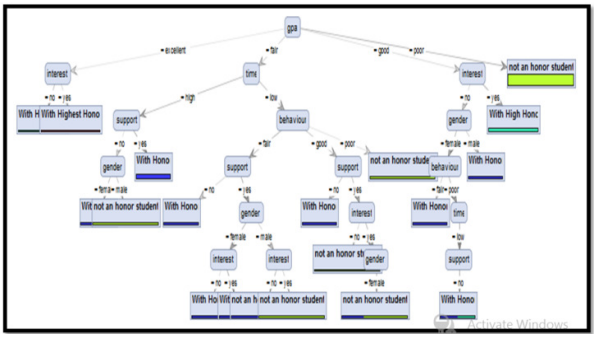


Fig 4 Obtained Decision Tree Model

IV. CONCLUSION

Predicting students' performance is mostly useful to help the educators and learners improve their learning and teaching process. This paper has reviewed previous studies on predicting students' performance with various analytical methods and thus used for preference in analyzing the methods and parameters that we will use in predicting the probability of a student in becoming an honor student. Thus the researchers have used students' grade point average (GPA), Psychometric factors and Students demographic as data sets based on the reviews.

Based on the results from the testing and evaluation, the researchers found out that using the Decision Tree Algorithm is efficient for predicting student performance for it has an accuracy of 90.67%. In conclusion, the meta-analysis on predicting students' performance has motivated us to carry out further research to be applied in our environment. It will help the educational system to monitor the students' performance in a systematic way and taking appropriate action for the students.

REFERENCES

[1] Bekele, R., Menzel, W. "A Bayesian Approach to Predict Performance of a Student (BAPPS): A Case with Ethiopian Students". Journal of Information Science (2013).

[2] Bhardwaj, K., Pal, S "Data Mining: A prediction for performance improvement using classification". International Journal of Computer Science and Information Security.Volume 9(4). (2011).

[3] Council N. "Knowing What Student Knows. The Science and Design of Educational Assessment".National Academic Press. Washington, D.C. 2001

[4] Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.

[5] G. Gray, C. McGuinness, P. Owende, An Application of Classification Models to Predict Learner Progression in Tertiary Education, in: Advance Computing Conference (IACC), 2014 IEEE International, IEEE, 2014, pp. 549–554

[6] Pandey, Sharma (2013). A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction

[7] Quinito (2014). A Framework for Students Academic Performance Monitoring and Evaluation System (SAPMES) for Higher Education Institutions

[8] S. T. Jishan, R. I. Rashu, N. Haque, R. M. Rahman, Improving accuracy of students final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique, Decision Analytics 2 (1) (2015)

[9] Surjeet K, Yadav, Bharadwaj, B. Pal B." Data Mining Applications: A comparative Study for Predicting Student's performance." International journal of innovative technology & creative engineering.Volume 1(12). (2012).

[10] V. Ramesh, P. Parkavi, K. Ramar, Predicting student performance: a statistical and data mining approach, International Journal of Computer Applications 63 (8) (2013)

[11] AmirahMuhammedShahiri, A Review on Predicting Student's Performance Using Data Mining Techniques (2013)