# IMPLEMENTING DATA MINING TOOL FOR STUDENTS' ACADEMIC PERFORMANCE PREDICTION

**Oladipo, F. O.**[*1], **Jokotoye, M. D.**[1]

[1]*Department of Computer Science, Federal University Lokoja, Kogi State, Nigeria.*

[*]***E-mail: francisca.oladipo@fulokoja.edu.ng***

## ABSTRACT

*This work describes the development of EmfactorPredictz, a tool developed for students' academic performance prediction. The tool applies a Data Mining model which comprises six algorithms (K-Nearest Neighbor, Naive Bayes, Decision Tree, Support Vector Machine, Linear Discriminant Analysis and Logistic regression) and taking into consideration the various factors that can affect students' academic performance, the tool is able to predict either the student is going to pass or fail with a degree of accuracy. Results gotten from this research showed that you can actually predict the performance of a student provided you can get the results he/she obtained from past evaluations. The research also revealed that other factors apart from grades can also affect the performance of students.*

***Keywords****: Data Mining, K-Nearest Neighbor, Linear Regression, Linear Discriminant Analysis, Decision Tree, Support Vector Machine, Naïve Bayes, Logistic Regression*

## 1.0 INTRODUCTION

As students are admitted into the university yearly, the amount of data collected about the admitted data has become enormous over the years. Although, universities collect an enormous amount of data about admitted students, this data does not fulfill its purpose of decision making that can improve the performance of students because no sense can be made from the data by mere looking at it (Ahmad, *et al*., 2015), hence the application of an automated method (Data Mining) that can make sense out of that large amount of data has to be adopted.

While Data Mining (DM) is generally defined as the process of discovering sensible patterns in a large chunk of data. The application of Data Mining to educational datasets, it is called Educational Data Mining (EDM) (Kabakchieva, 2012). In addition, (Dutt, *et al*., 2017) gave a definition of EDM according to the International Educational Data Mining Society as "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.”

Universities stand a higher chance of providing quality education (which is their primary aim) if they better understand their students and also predict students' performance accurately as

***Corresponding Author**

**How to cite this paper:** Oladipo, F. O., Jokotoye, M. D.(2018). Implementing data mining tool for students' academic performance prediction. *Confluence Journal of Pure and Applied Sciences (CJPAS), 2*(1), 226 - 234.

results of such prediction will help admission officers classify students into suitable (likely to do well) or unsuitable (not likely to do well) and then provide support like moral assistance, tutoring resources etc. for the unsuitable students to help them throughout their stay in the university (Ogwoka, *et al.*,). Consequently, from the prediction, students would be able to identify their weaknesses beforehand and can as a result improve themselves adequately by developing a suitable learning strategy and also, lecturers will be able to plan their lectures strategically as per the need of students so as to help them in their learning (Ani *et al.*, 2015).

This work will study the approaches that various researchers have adopted for the prediction of students' academic performance, then use those approaches to develop a hybrid and more sophisticated approach to predicting students' academic performance. We developed a tool that predicts students' academic performance using this approach. The tool applies a Data Mining model as its back end to enable the prediction. The tool serves as an Early Warning System for students and also helps universities better understand their students.

### 1.1 Approaches to Predicting Academic Performance

Predicting students' academic performance can be a tricky thing to do as various factors and conditions have to be considered. Over the years, various researchers have approached the problem of predicting students' academic performance in many different ways but classifying them into two gives us the following:

(A) Finding Dominant Factors

This approach focuses on the important factors ranging from physical, environmental, psychological, psychosocial etc. which may affect students' academic performance. The degree at which these factors affect students' performance is then calculated and the most dominant of these factors is used as the basis for the prediction of the academic performance through the application of various data mining techniques. This approach is the most popular as various researchers including (Fayombo, 2012; Kplolovie, *et al.*,2014; Kovačić, 2010) and several others adopted it for their research.

(B) Finding Dominant Subject/Course

This approach uses the student's performance in one key course as the basis for his/her academic performance. For example; Mathematics is seen as the building block of Computer science as English Language is seen as the building block of Journalism. So, in order to predict the students' academic performance in that field of study, his/her performance in those key courses will most likely be the determinant for the students' success or failure in the field. This approach is the least popular as there aren't many situations in which the performance in one subject determines the overall performance of the student.

### 1.2 Review of Related Work

Several researches based on different approaches abound on the subject of students' academic performance prediction. (Verma*et al.*, 2006) focused on the implementation of data mining techniques and methods for acquiring new knowledge from data collected by universities. The aim of the research work was to find out if there are any patterns in the available data that could be useful for predicting students' performance at the university based on their personal and pre university characteristics and also to find out which features in the data are the strongest predictors of university performance. The research work used a software tool called WEKA together with Popular WEKA classifiers including a common decision tree algorithm C4.5 (J48), two Bayesian classifiers (NaiveBayes and BayesNet), a Nearest Neighbour algorithm (IBk) and two rule learners (OneR and JRip).

Using data mining processes, particularly classification, (Al-Radaideh*et al.*,2006) looked into enhance the quality of the higher educational system by evaluating student data and highlighting the main attributes that may affect the student performance in courses. The CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) was used to develop an effective model and the WEKA toolkit was used to calculate the model's accuracy.

Olani (2009) deployed a number of prior academic achievement measures (including preparatory school grade average point (GPA), aptitude test scores, university entrance exam scores) and some psychological variables (achievement motivation and academic self-efficacy). Prior academic achievement records of 3301 first year university students were obtained from archival sources and 214 out of the 3301 students also filled in a self-report which contains information about some psychological variables. The data was analyzed using standard multiple regression analysis and stepwise multiple regression analysis. The result revealed that prior academic achievement measures and psychological variables in combination accounted for 17% of the variance in students' university GPA scores.

Erimafa*et al.*, (2009) deployed discriminant analysis to predict the class of degree obtainable in a university system. The data collected for this study were from student's academic records for 100 level and 200 levels, in the Department of Statistics, from 2004 to 2007 academic session in a University. The result of the study successfully predicted the classes of degree of 87.5% of graduating students of the University students.

Similarly, a model for student's academic performance prediction based on a number of influencing factors was developed by (Affendey*et al.*, 2010). The WEKA model was deployed for the analysis of the dataset collected and the result obtained was used for the student's academic performance prediction. The dataset used comprised of 2,427 number of student record and 356 attributes of students registered between 2000-2006. The data was applied to three classifiers (Naïve Bayes, tree and function classifiers). The classifiers categorized data into either First class - Second class upper or Second class - Third class lower classes. A 10-fold cross validation was used to get the accuracy of the classifiers.

While (Fayombo, 2012; Kovačić, 2010) highlighted the extent to which socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course programme and course block) may influence the persistence or dropout of students at the Open Polytechnic of New Zealand, (Oyelade 2010) developed a system that successfully clusters students into categories based on academic performance. Presenting K-means clustering algorithm as a simple and efficient tool to monitor the progression of students' performance in higher institution, their work used k-means clustering algorithm for analyzing students' results using Grade Point Average (GPA) as the indicator of students' performance. A similar approach by (Shovon and Haque, 2012) was aimed at devising a way of improving the academic performance of students using K-means clustering algorithm and Decision tree. From the 50 training samples gotten for the research, the clustering algorithm clustered the students into three categories namely; High, Medium and low with a student in the high cluster having CGPA >= 3.50, a student in the medium cluster having CGPA between 2.20 and less than 3.50 and finally, a student in the low cluster having a CGPA of <= 2.20. The Decision tree algorithm was used to tell which students are in need of help from the instructor.

Garc´ıa-Saiz and Zorrilla (2011) compared the performance and interpretation level of the output of different classification techniques applied on educational datasets and propose a meta-algorithm to preprocess the datasets and improve the accuracy of the model. The result of the experiment revealed that there is not one algorithm that obtains a significantly better classification accuracy than the others. In fact, the accuracy depends on the sample size and the type of attribute.

Applying Bayesian classification algorithm, (Pandey, 2011) successfully classified students division based on the previous year database. Performed at VBS Purvanchal University, data was gathered from different degree colleges affiliated with Dr. R. M. L. Awadh University, Faizabad, India to aid the work.

Using WEKA tools, a study of the Naïve Bayes, Decision tree and neural networks and how they perform in the prediction of student's academic success was conducted by (Osmanbegović and Suljić, 2012). The researchers also studied in addition, the degree to which various input factors affect the success of a particular student. The result of the experiment showed that Naïve Bayes outperforms Decision tree and Neural networks in terms of predictive accuracy.

With a view to reveal the high potential of data mining applications for university management, Kabakchieva (2012) developed a student performance prediction model using WEKA and its various classifiers including a rule learner (OneR), a common decision tree algorithm C4.5 (J48), a neural network (MultiLayer Perceptron), and a Nearest Neighbor algorithm (IBk). The performed research was based on the CRISP-DM (Cross Industry Standard Process for Data Mining) model. The model was developed at one of the famous Bulgarian universities.

The idea of using temporal association mining to predict a student's future academic performance was worked on by Basha*et al*., (2012). The research work involved identifying existing patterns in the historic data, and saving it. Then through the comparison of the current performance of a student with the existing patterns, the possible performance of the student in the future can be predicted. In the process, any new pattern that is discovered is identified. Student data set from 294 affiliated colleges of Kakatiya University were collected from 2002 to 2007 which contains results and marks for B.Sc.(M), B.Sc. (B), B. Com and B.A. courses from these colleges. There are approximately 500,000 records in the dataset.

Li *et al*., (2012) assessed students' performance in Elements of Statistics using data from UWF (University of West Florida) for 2008, 2009 and 2010 fall semesters. To access the students' performance, they found the relationship between the students' performance and other performance related factors including: college and high school GPA (grade point averages), prerequisite algebra courses, and scores on standardized examinations. Results revealed that the student GPA is the most reliable predictor of students' performance in Elements of Statistics.

Tekin (2014) looked to predicting a

students' GPA at graduation early using data mining techniques. Neural Networks (NN), Extreme Learning Methods (ELM) and Support Vector Machine (SVM) were applied to data of computer education and instructional technology students to predict their GPAs at graduation. The results showed that SVM was the most successful in prediction with an accuracy of 97.8%, then ELM with 94.94% and the least accurate was NN with a 93.76% accuracy.

Chen *et al*., (2014) approached the problem of predicting student's academic performance based on the artificial neural network (ANN) with the two meta-heuristic algorithms inspired by cuckoo birds and their lifestyle, namely, Cuckoo Search (CS) and Cuckoo Optimization Algorithm (COA). The standard CS and standard COA were separately utilized to train the feed-forward network for prediction. The result of the study demonstrated that both CS and COA can be used to train ANN and the ANN trained by COA obtained slightly better results for predicting student academic performance in this study.

Predicting students' academic performance requires vast knowledge in Machine learning and Data mining. Several researchers have used different approaches to develop models that predict the academic performance of students. The fact that only models exist without any system that they are integrated into gives justification to this research work.

## 2.0 MATERIALS AND METHODS

This research is in two parts: the development of a backend in the form of a machine learning model that predicts students' academic performance by applying six algorithms (KNN, Naive Bayes, Decision Tree, Support Vector Machine, Linear Discriminant Analysis and Logistic regression) over data extracted from Educational, Environmental, Social, Family related and personal features to develop the prediction model. And the development of a frontend system that interfaces with this model using the Knowledge Discovery in Databases (KDD) methodology. The Design artifacts were developed to extract/ discover knowledge from data as labeled with the characteristics above.

### 2.1 Data Mining Models

Classification and regression are two important DM goals that are carried out under supervised learning, where a model is adjusted to a dataset made up of k ∈ {1, ..., N} examples, each mapping an input vector (xk1, . . . ,xkI ) to a given target yk. The main difference between them is in terms of the output representation, (i.e. discrete for classification and continuous for regression). Their evaluation also differs because in classification, models are often evaluated using the Percentage of Correct Classifications (PCC), while in regression the Root Mean Squared Error is used. Both are represented mathematically below:

$$\mathbf{PCC} = \sum_{i}^{n} \frac{\phi \times 100(\%)}{N} \qquad \text{(i)}$$

Where $\emptyset = 1$ if $y_i = y$ and 0 otherwise

$$\mathbf{RMSE} = \sqrt{\sum_{i-1}^{n}(y_i - y)^2/N} \qquad \text{(ii)}$$

### 2.2 Student Data

The data collected for this research was from the UC Irvine (UCI) Machine learning repository. The data is named Student performance data. The data contains 395 rows and 33 columns. A description of the labeled elements in students is given in Table I.

Table I: Description of the student data fine - tuned for the research

| Field | Description |
|---|---|
| Sex | The Gender of the student |
| Age | The Age of the student |
| Address | The address of the student |
| Famsize | The number of people in the students' family |
| Pstatus | The marital status of the students' parent |
| Medu | The education status of the students' mother |
| Fedu | The education status of the students' father |
| Mjob | The Job of the students' mother |
| Fjob | The Job of the students' father |
| Reason | The student's reason for joining the school |
| Guardian | The guardian of the student |
| Traveltime | The travel time of the student |
| Studytime | The study time of the student |
| Failures | The failure rate of the student |
| Scholarship | If the student is on scholarship or not |
| Course | If the student is happy with the course of study or not |
| Easy | If it is easy to pay the school fees or not |
| EActivities | If the student is involved in extra curriculum activities |
| SActivities | If the student is involved in sporting activities or not |
| Further | If the student wants to further his her education |
| Internet | If the student has internet connectivity at home or not |
| Romantic | If the student is in a romantic relationship or not |
| Famrel | The degree of the student's family relationship |
| Freetime | The student's freetime |
| Gout | How often the student goes out |
| Dalc | The student's daily consumption of alcohol |
| Walc | The student's weekly consumption of alcohol |
| Health | The student's current health status |
| Absences | The percentage of times the student was absent from school |
| G1 | The student's grade |
| G2 | The student's grade |

## 2.3 High Level Model

The High Level model of *EmfactorPredictz* (Fig. 1) presents a simplified representation of the system. The system consists of two major parts which are the information management system and the prediction management system. The information management system takes care of users' registration, login, and other information related process; while the prediction management system takes care of predicting students' academic performance. It is also where the contents of the system is managed. It serves as the system control.

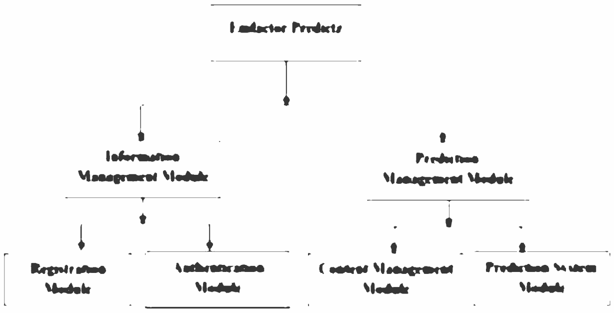

Emfactor Predictz
Information Management Module
Prediction Management Module
Registration Module
Authentication Module
Content Management Module
Prediction System Module

Fig. 1. High Level Model of *EmfactorPredictz*

## 2.4 Data Flow Diagram

The data flow diagram for the *EmfactorPredictz*system illustrated by (Fig. 2) shows the user of the system. When a user goes to the homepage of the web application, he she is required to login before gaining access to the system. If the user is not registered, he she is required to register after which the user will have access to make prediction and view the prediction result.

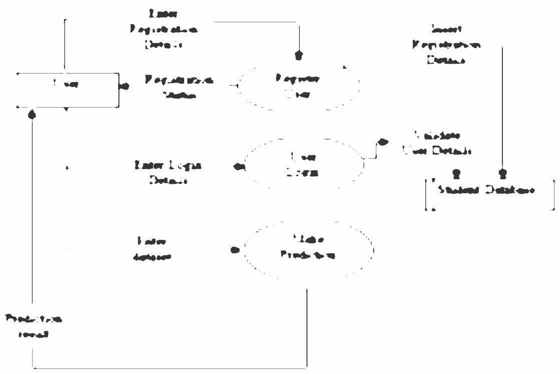


Fig. 2. Data Flow Diagram of *EmfactorPredictz*

## 3.0 RESULTS AND DISCUSSION

This work has been able to analyze and predict students' performance through the application of selected Machine Learning algorithms to a dataset of labelled corpus from the UCI Repository and the Development of System Architecture and Design Artefacts in the form of system architecture (Fig. 3); Use Case Description (Fig. 4) and Activity Diagram (Fig.

5).

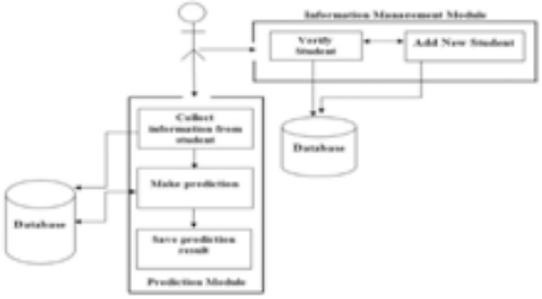


Fig. 3. System Architecture

The use case shows the various actions that can be performed by the user on the system andthe activity diagram shows the control flow.

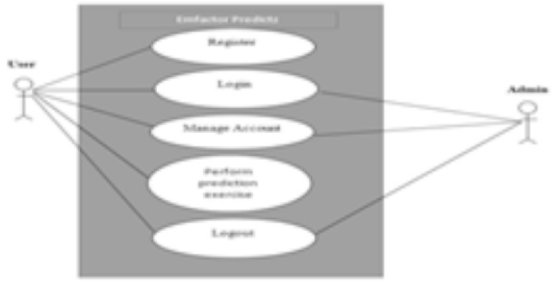


Fig.4. Use Case Diagram of *EmfactorPredictz*

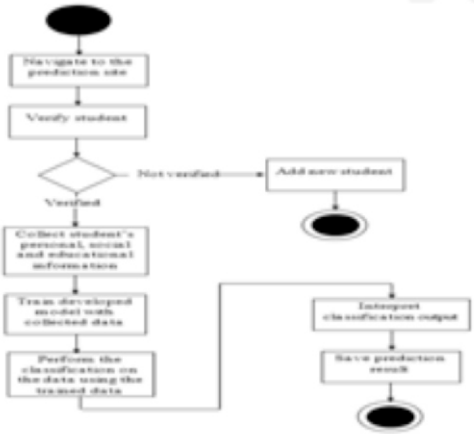
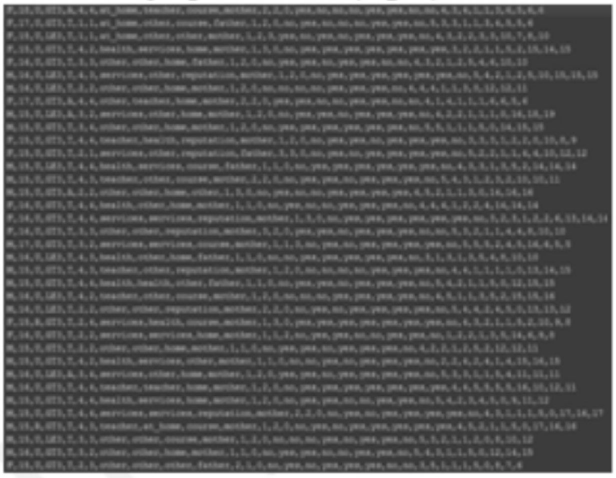

Navigate to the prediction site
Verify student
Not verified
Add new student
Verified
Collect student's personal, social and educational information
Train developed model with collected data
Perform the classification on the data using the trained data
Interpret classification output
Save prediction result

Fig.5. Activity Diagram

Qualitative and quantitative research methods were used in this research. By preparing the data obtained from the UC Irvine (UCI) Machine Learning repository (Fig.6) using feature selection -using Pearson's correlation coefficient, we were able to get a structured and noise free dataset. Algorithms such as KNN, Naive Bayes, Decision Tree, Support Vector Machine, Linear Discriminant Analysis and Logistic regression were used to develop this prediction model. The backend model works by choosing one algorithm from the list of algorithms based on K-Fold cross validation technique, and then applies the best algorithm for that particular test to carry out the prediction. Based on past training and testing carried out (Fig.7), it has predicted students' academic performance with a very high degree of accuracy up to 92.5%, (Fig. 8).



```
[illegible]
```

Fig.6. Dataset which serves as input to the model

```
[illegible]
```

Fig.7. Output of the Data Mining Model

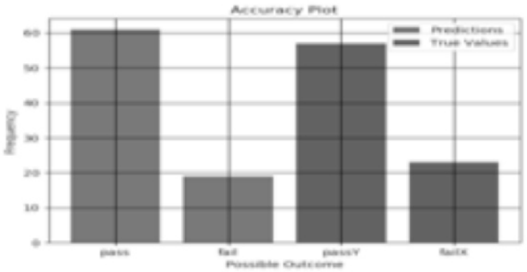


Fig. 8. Accuracy Plot of the Data Mining Model

## 4.0CONCLUSION

In an attempt at solving the problem of predicting students' academic performance, a Data Mining (DM) model integrating six algorithms ;K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Discriminant Analysis, Logistic Regression and Decision Tree was defined. Using the dataset from UC Irvine (UCI) Machine Learning repository, the model drops fields that are not useful using Pearson's correlation coefficient and selects the algorithm to be used for prediction by testing the accuracy of each algorithm using 10-fold cross validation. The model classifies students into two classes (pass and fail). A web application with a friendly user interface was then developed to serve as a front end for the developed DM model.

## REFERENCES

Ahmad, F., Ismail, N. H., and Aziz, A. A. (2015). “The Prediction of Students' Academic Performance”. Applied Mathematical Sciences, 12.

Affendey, L., Paris, I., Mustapha, N., Sulaiman, M. N., and Muda, Z. (2010). Ranking of influencing factors in predicting of Students' academic perfoemance. Information Technology Journal, 832-837.

Al-Radaideh, Q. A., Al-Shawakfa, E. M., and Al-Najjar, M. I. (2006). Mining Student Data Using Decision Trees. The 2006 International Arab Conference on Information Technology, (pp. 1-5).

Ani, A., Anil, M., and Manisha. (2015). Performance Analysis and Prediction in Educational. International Journal of Computer Applications.

Basha, S. A., Kumar, Y. R., Govardhan, A., and Ahmed, M. Z. (2012). Predicting Student Academic Performance Using Temporal Association Mining. International Journal of Information Science and Education, 21-41.

Chen, J.-F., Hsieh, H.-N., and Do, Q. H. (2014). Predicting Student Academic Performance: A Comparison ofTwo Meta-Heuristic Algorithms Inspired by Cuckoo Birds forTraining Neural Networks. Algorithms(7), 538-553.

Dutt, A., Ismail, M. A., and Herawan, T. (2017). A Systematic Review on Educational Data. IEEE Access, 5(15991), 16.

Erimafa J.T., Iduseri A. and Edokpa I.W (2009). Application of discriminant analysis to predict the class of degree for graduating students in a university system. International Journal of Physical Sciences, 16-21.

Fayombo, G. A. (2012). Relating emotional intelligence to academic achievement among university students in Barbados. The International Journal of Emotional Education, 43-54.

Garc´ıa-Saiz, D., and Zorrilla, M. (2011). Comparing classification methods for predicting distance students' performance. JMLR: Workshop and Conference Proceedings 17, 26-32.

Kabakchieva, D. (2012). Student Performance Prediction by Using Data. International Journal of Computer Science and Management Research.

Kovačić, Z. J. (2010). Early Prediction of Student Success: Mining Students enrollment data. Proceedings of Informing Science & IT Education Conference, (pp. 1-19).

Kplolovie, P. J., Joe, A. I., and Okoto, T. (2014). Academic Achievement Prediction: Role of Interest in Learning and Attitude towards school. International Journal of Humanities Social Sciences and Education (IJHSSE), 1(11), 73-100.

Li, K., Uvah, J., and Amin, R. (2012). Predicting Students' Performance in Elements of Statistics. US-China Education Review, 875-884.

Ogwoka, T. M., Cheruiyot, W., and Okeyo, G. (2015). “A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms”. International Journal of Computer Applications Technology and Research, 693-697.

Olani, A. (2009). Predicting First Year University Students' Academic success. Electronic Journal of Research in Educational Psychology, 7(3), 1053-1072.

Osmanbegović, E., and Suljić, M. (2012). Data Mining Approach for Predicting Student performance. Journal of Economics and Business, X(1), 1-11.

Oyelade, O. J., Oladipupo, O. O., and Obagbuwa, I. C. (2010). Application of k-Means Clustering algorithm for. (IJCSIS) International Journal of Computer Science and Information Security, 4.

Pandey, U. K. (2011). Data Mining : A prediction of performer or underperformer using classification. International Journal of Computer Science and Information Technologies, 686-690.

Shovon, M. H., and Haque, M. (2012). An Approach of Improving Student's Academic performance by using K-means clustering algorithm and Decision tree. International Journal of Advanced Computer Science and Applications,, 146-149.

Tekin, A. (2014). Early prediction of students' grade point averages at graduation. Eurasian Journal of Educational Research, 207-226.

Verma, K., Singh, A., and Verma, P. (2006). A Review on Predicting Student Performance. International Journal of Current Engineering and Scientific Research, 6.