Matrix Factorisation for Predicting Student Performance

Edgar Jembere, Randhir Rawatlal, Anban W. Pillay
College of Agriculture, Engineering and Science
University of KwaZulu-Natal
Durban, South Africa

***Abstract*—Predicting student performance in tertiary institutions has potential to improve curriculum advice given to students, the planning of interventions for academic support and monitoring and curriculum design. The *student performance prediction* problem, as defined in this study, is the prediction of a student's mark for a module, given the student's performance in previously attempted modules. The prediction problem is amenable to machine learning techniques, provided that sufficient data is available for analysis. This work reports on a study undertaken at the College of Agriculture, Engineering and Science at University of KwaZulu-Natal that investigates the efficacy of Matrix Factorization as a technique for solving the prediction problem. The study uses Singular Value Decomposition (SVD), a Matrix Factorization technique that has been successfully used in recommender systems. The performance of the technique was benchmarked against the use of student and course average marks as predictors of performance. The results obtained suggests that Matrix Factorization performs better than both benchmarks.**

Keywords—component; Matrix Factorisation; Singular Value Decomposition; Educational Data Mining; Student Performance Prediction

I. INTRODUCTION

There is growing interest in evidence-based, data driven approaches to decision-making in educational institutions. This interest is driven by the large stores of electronic data collected by most modern institutions and a thriving research community that seeks to apply data mining and machine learning techniques in Education. The Educational Data Mining [1,2] community investigates techniques that can be used to effectively glean potentially actionable knowledge from Educational Databases with the aim to better understand students and the settings they learn in.

An important area of Educational Data Mining studies the development of performance prediction tools that will aid faculty advisors to provide evidence based guidance to students on module/course choices; help to determine bottlenecks in the system that will provide evidence based support in decision making for curriculum reform and design, feed into academic monitoring and support efforts and inform policy decisions such as enrollment criteria.

Previous approaches to the prediction problem have differed along two dimensions: the data used and the machine learning/data mining techniques employed. Several approaches have used demographic and socioeconomic data in contrast to approaches that use only marks obtained in previous modules [11].

Several machine learning techniques have been explored for educational data mining with supervised learning approaches being shown to be best suited to the prediction problem [11]. In these approaches, a learning algorithm builds and validates a model after training on a dataset that can then be used to make predictions. Larger training datasets often yield models with higher predictive power that can generalize well. Classical supervised learning techniques include: Naïve Bayes, Bayesian Belief Networks, Logistic Regression, Support Vector Machines, Artificial Neural Networks and Decision trees—all of which have been shown to be useful in solving the prediction problem [11].

In this research, Matrix Factorization (MF) [3], a technique that has been shown to be effective in recommender systems, is applied to the student performance prediction problem. The technique uses only the past performance of students, i.e. no demographic or socio-economic data is required. We hypothesize that the various factors influencing student performance are hidden in the marks that students have obtained in past modules and that these marks are sufficient to predict a student's marks. This hypothesis is generally true in recommender systems, where it has been shown that the variability in the ratings that users assign to items is sufficient for predicting the ratings of unrated items. The performance of the Matrix Factorization (MF) as a prediction tool is compared to the use of the student's average mark and the course average mark as predictors of future performance.

The rest of this paper is organized as follows: Section II gives an overview of Matrix Factorization. In Section III we discuss related work on the student performance prediction problem in general and on the application of Matrix Factorization for Student Performance prediction in particular. The research methods and the nature of the data set used in this study are discussed in Section IV. Section V discusses how the parameters used for the MF model were determined and in Section VI we give the results of the study. Section VII discusses the findings of our study and Section VIII concludes the paper.

II. MATRIX FACTORIZATION

In this section, a brief overview Matrix Factorization (MF) and a motivation for applying it to the student performance prediction problem is given. The use of MF-based approaches

for grade prediction is postulated on the belief that there is a low-dimensional latent feature space that can jointly represent both students and courses. The latent space can correspond to the space of knowledge components that can explain variability in student marks.

In this study, we used Singular Value Decomposition (SVD) [4, 5]. SVD is a well-known linear algebra technique that decomposes/factorizes an arbitrary matrix, A into a product of three matrices U, Σ, and V, as shown in Equation 1

$$A = U \Sigma V^T \quad (1)$$

where Σ is a $r \times r$ diagonal matrix, and r is the rank of the matrix A. In recommender systems, the matrix A is a $n \times m$ user-item matrix whose cells hold each user's rating for a given item. The matrix U is a mapping of users to the hidden features that explain variability in the data, and the matrix V^T is the interaction of the items and the hidden factors. The rank of a matrix is the number of linearly independent rows or columns in the matrix. Typically, the first $k < r$ features are picked to explain variability in the ratings data. The matrix A is generally a sparse matrix, and the goal of SVD is to predict the rating that a given user will give to an item she has not yet rated.

In this study, the user and item concepts in recommender systems is mapped to the student and course concepts. The mark a student obtained for a given module is equivalent to the user's rating of an item. Each course j is represented by a feature vector v_j and each user i by a feature vector u_i. The predicted rating for user i on item j is a dot product of the user vector u_i and item vector v_j given by Equation 2.

$$r_{ij} = u_j^T . v_i \quad (2)$$

The major difference between the user-item rating matrix and the student-course matrix is that the range of the ratings is usually [1, 5] while the range of student marks is [0, 100].

III. Related Work

A recent survey on the performance prediction problem [11] suggests that grade point average and internal assessments (tests and assignments) are important attributes for predicting performance. The authors conclude that external assessments (marks obtained in the final exam) play an important role in prediction accuracy. The survey also shows that supervised learning approaches are best suited to the problem.

A recent study [12] used machine learning to determine the factors that influence academic performance. They conclude that lecture attendance, lecturer competency, school leaving results and type of high school were important factors. They found that the Support Vector machine produced the highest accuracy. Their work, however, predicted only performance in the first year of university. A similar study [13] also attempted to predict performance of student in their first year to classify students at high, medium or low risk of failure. The authors used several factors pertaining to personal history of the student, their behavior and their perceptions. They concluded that prediction rates using several machine learning techniques were not good.

Matrix Factorization is one of many Collaborative filtering techniques that has been applied it to the student performance prediction problem. Examples of such work include [1, 6, 7, 8].

The work by [6] was the first to apply Matrix Factorization on the KDD Challenge 2010 dataset. This data was collected from a computer-aided-tutoring system and the authors predict whether a student will successfully complete a given step on the first attempt. In the training set this information is encoded as a 1 if a student successfully completed a given step on the first attempt and 0 otherwise. Predictions were then made in the range [0, 1] and the Root Mean Square Error was calculated accordingly. Matrix Factorization was found to outperform Regression Models and Classical user-item collaborative filtering. The paper also showed that a combination of MF and classical user-item collaborative filtering has the potential to improve prediction accuracy. The work by [1] was also done on the same dataset. The results from [1] were comparable to that of [6] with a RMSE of approximately 0.3 in the prediction range [0, 1].

The work reported in [7] extended Matrix Factorization with clustering in order to exploit the local geometric structure in the data. This was achieved by identifying groups of similar students and tasks based on the corresponding skill profiles. This work also used the KDD Challenge 2010 dataset to test their solution.

The work most related to the research reported in this work is by Polyzou and Karypis [8]. They used the data from Computer Science and Engineering (CSE) and Electrical and Computer Engineering (ECE) programs at the University of Minnesota from Fall of 2002 to Spring of 2014. However, the data was converted into the GPA range [0, 4] and did not use raw student marks.

The study reported in this paper differs from previous work in that we use the actual marks obtained in previous courses and we aim to predict the actual mark a student will obtain. The range of marks is thus [0, 100].

IV. Methods and Datasets

The goal of this research was to show that MF can be used to solve the student performance prediction problem. The Matrix Factorization solution was benchmarked against the student average mark and the course average mark as predictors of future performance. Comparison with other machine learning technique will be considered for future work. Unlike other work that used Matrix Factorization for the same problem, this work used the actual marks that the student obtained rather than a mapping to grades or ranges. Matrix Factorization was applied on this data to get the hidden interaction factors that will be used for predicting student performance.

The data was obtained from the BSCSIT programme at the University of KwaZulu-Natal. The data includes the students' marks for the 2013-2014 cohorts. Only students who completed the programme were considered. The data was very sparse, because students have many elective modules to choose from. The data had 501 students and 137 modules. Owing to the sparsity of the data and the fact that we wanted the student marks to remain in the range [0, 100], no normalization of the data was done.

The Matrix Factorization technique, Singular Value Decomposition (SVD) [4, 5 9], was used in the study. For parameter estimation, the predictions from MF were used without the user and item bias correction. After fine tuning the parameters, experiments were then conducted, with MF with student and course bias correction. The bias correction was done as shown in Equation 3

$$\widehat{r_{ij}} = w_s s_i + w_c c_j + w_{mf} c_j^T . s_i \quad (3)$$

where w_s , w_c, and w_{mf} are weights which sum up to 1. The student's average mark was used to estimate the student bias and the course average mark was used to estimate the course bias. This approach yielded better results than effecting the bias correction through subtractive normalization.

The MF solution was implemented in Java. For each run, the program sampled a given number of entries from the entire data set to form the test set. The sampled cases were then removed from the data and the cases that remain formed the training set. An MF model was then built from the training set and used on the test set. The average Root Mean Square Error (RMSE) for each run was computed and written to a file. For each run, a test set was sampled, and the corresponding average RMSEs for MF, the student average and the course average were recorded. Due to RMSEs being recorded for a given sample for each run, the statistical test for difference between two means in paired samples was used to ascertain whether MF significantly performs better than the student and the course average marks.

V. Parameter Estimation

All experiments for determining optimal parameters were performed with Matrix Factorization predictions without correction for user and course bias.

A. Determining the learning rate(α)

To determine the optimal learning rate (α), the number of factors and the regularization parameter(λ) were set to 5 and 0.1 respectively. For each learning rate value, 10 runs were conducted and the RMSEs were recorded. Learning rates greater than 0.0003 were found to return no predictions. As the learning rate approached 0.0003 the RMSE for the prediction increased. A learning rate of 0.00004 was found to give optimal RMSE values.

B. Determining the optimal number of hidden features

The optimal learning rate (α=0.0004) from the previous experiment was used to determine the number of hidden factors needed for optimal student performance prediction. The regularization parameter (λ) was set to 0.1. Figure 1 shows that the first factor explains most of the variability in the student marks. The RMSE from a model with one factor is comparable to that with two and three factors. For more than three factors, MF begins to add unnecessary noise to the predictions. We therefore settled on a model with three hidden factors.

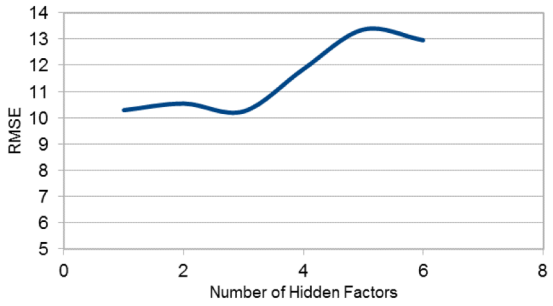

Figure 1: Hidden Factors

C. Determining the regularisation parameter (λ)

To determine the regularization parameter (λ), the learning rate was set to 0.00004 and the number of factors to 3. As shown in Figure 2, the regularization parameter was found not to have a significant influence on the RMSE for MF. We settled on using $\lambda = 1$ for all experiments

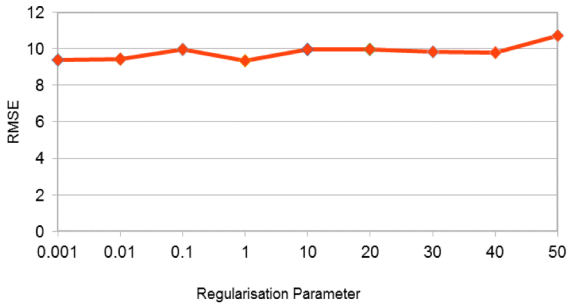

Figure 2: Determining the Regularization Parameter (λ)

VI. Evaluation of Matrix Factorisation for Student Performance Prediction

In evaluating the use of Matrix Factorization for student performance prediction we chose to benchmark it against the use of the student's average mark and the course average mark as predictors of student performance. Matrix Factorization should at least do better than these two benchmarks.

Since there seem to be no significant difference in the RMSE across different regularization parameters, this experiment was done with varying regularization parameters. 20 runs were done for each λ and the number of factors and the learning rate were kept at 3 and 0.00004 respectively. The results show that MF seems to perform better (although not statistically significantly better) than using the student average mark or the course average mark to predict a student's performance. Significant differences with the student average mark were only observed at 90% confidence level for $\lambda = 1$. In all other cases, pairwise statistical test for the difference in RMSE were inconclusive.

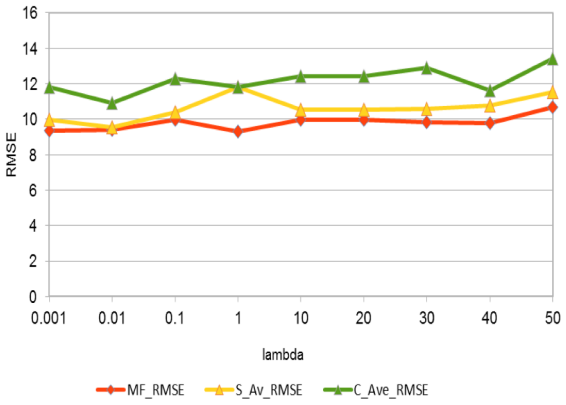

Figure 3: Comparison of MF with Student and Course Average Marks across different Lambdas

The box-plot in Figure 4 depicts the difference between the distribution of RMSE for the MF_RMSE, Sav_RMSE, and Cav_RMSE across different regularization parameters. The results show that MF seems to outperform the student and course average. The RMSE for MF was found to be approximately 9.7, compared to that of the student and class average, which were approximately 10.3 and 12.4 respectively.

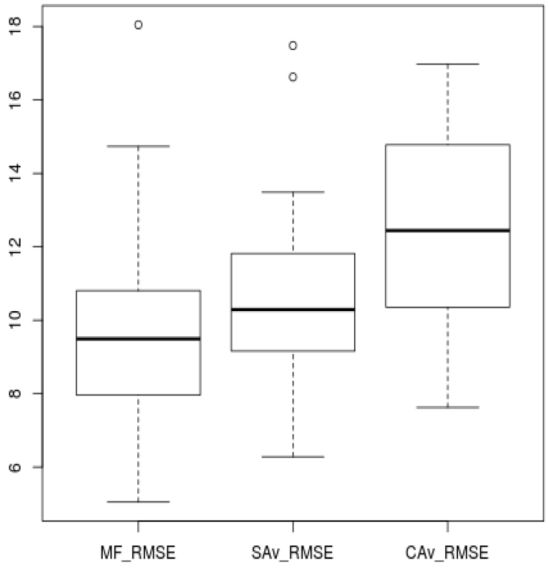

Figure 4: Comparison of MF with Student and Course

The experiment was performed again with λ =1. We increased the number of runs to 50 to check whether the test remains inconclusive at 0.05 level of significance. The statistical test showed pairwise significant difference in RMSE for all the pairs. MF outperformed student's average mark and the course average mark (See Table 1). The pairwise statistical test for difference between two means in paired samples showed significant statistical difference for both the pairs MF_RMSE and Sav_RMSE, and MF_RMSE and Cav_RMSE. Figure 5 shows the box and whisker plot for the distribution of RMSE across the different prediction techniques.

Table 1: T-Test for difference between two means on paired samples-50 runs

Paired Samples Test					
	Paired Differences		t	df	Sig.
	Mean	Std. Error			
MF_RMSE - SAv_RMSE	-1.184	.281937	-4.200	49	.000
MF_RMSE - CAv_RMSE	-2.535	.415702	-6.099	49	.000
SAv_RMSE - CAv_RMSE	-1.351	.476010	-2.839	49	.007

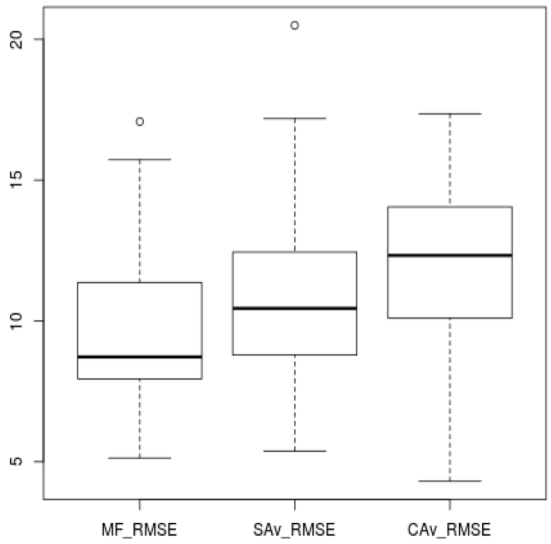

Figure 5: Comparison of MF with Student and Course Average Marks - 50 runs

A. Analysis of prediction error within the Grades

In this section, we investigate the performance of MF when marks are graded. The student marks were graded as shown in column 1 of Table 2. Table 2 summarizes the statistical tests that were carried with test datasets of sizes 1, 10, and 50 for every run. Figures 6-8 show how the RMSEs for MF, Student Average Mark and Course Average Mark differ for each grade. MF seems to perform better for marks in the ranges [70-79] and [80-100]. However, there was no statistically significant difference between MF and Student average mark when using test sets of sizes 1 and 10. Statistically significant differences were observed for the test set of size 50.

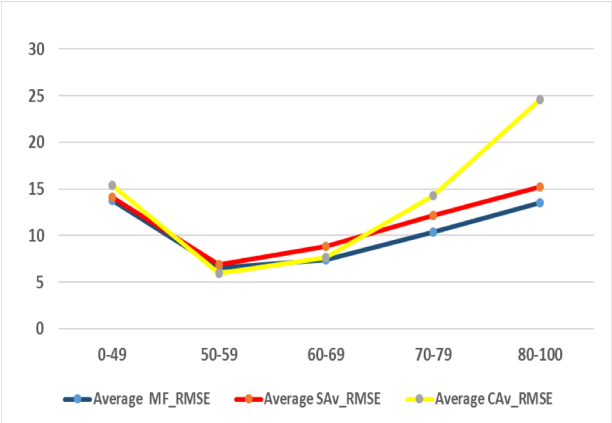

Figure 6: Comparison of MF to Student and Course Average, Test Data Size =10

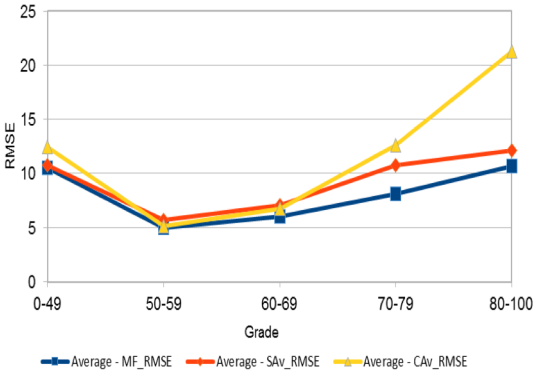

Figure 7: Comparison of MF to Student and Course Average, Test Data Size =1

VII. Discussion

The Matrix Factorization prediction with both user and course bias correction was generally found to perform better than using the student and course average mark. The RMSE for MF was 9.7 whilst that for the student average and the course average were approximately 10.3 and 12.4 respectively. However, the RMSE variedly widely depending on the actual grade a student obtained. Prediction of the marks within the range [50-69] were more accurate with RMSE in the range [5-7] for all prediction approaches. This is mainly because student marks are usually concentrated in this range, hence even the student and course average does well in predicting marks within this range. Marks within the ranges [70-79] and [80 - 100] were the most difficult to predict since few students get mark in this range. MF was seemingly performing better than the student and the course average in these ranges. However, the RMSE for MF was higher in this range than in the other ranges.

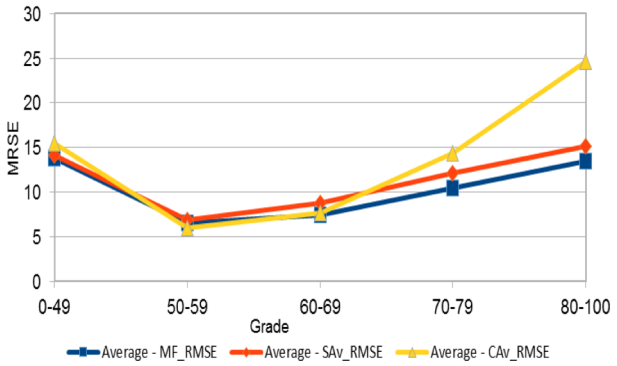

Figure 8: Comparison of MF to Student and Course Average, Test Data Size =50

Table 2: Paired Samples Test for Difference between two means on the different approaches to predicting Student marks

Grade		Test Data Size		
		10	1	50
0-49	MF_RMSE-SAv_RMSE	.233	.722	.227
	MF_RMSE-CAv_RMSE	**.007**	**.016**	**.000**
	SAv_RMSE-CAv_RMSE	.128	.091	**.000**
50-59	MF_RMSE- SAv_RMSE	.904	.074	.142
	MF_RMSE- CAv_RMSE	.645	.834	.059
	SAv_RMSE- CAv_RMSE	.753	.426	.006
60-69	MF_RMSE- SAv_RMSE	.063	.072	**.000**
	MF_RMSE- CAv_RMSE	.714	.258	.475
	SAv_RMSE- CAv_RMSE	.121	.656	**.001**
70-79	MF_RMSE- CAv_RMSE	.363	**.003**	**.015**
	MF_RMSE- CAv_RMSE	**.000**	**.002**	**.000**
	SAv_RMSE- CAv_RMSE	**.011**	.280	**.003**
80-100	MF_RMSE- CAv_RMSE	.187	.307	**.001**
	MF_RMSE- CAv_RMSE	**.000**	**.000**	**.000**
	SAv_RMSE- CAv_RMSE	**.000**	**.002**	**.000**

VIII. Conclusion

Matrix Factorization was applied to the student performance prediction problem. The results of our study showed that MF can be used for predicting student marks. Its suitability for this problem stems from the fact that it is robust with sparse data. We postulate that the performance of Matrix Factorisation could be

improved if more data was available. The dataset used had only 501 students. This is significantly smaller than data used in typical recommender systems. The number of students was too few for effective extraction of latent features that can concretely explain variability in student marks.

As part of our future work, we will explore mechanisms for improving the predictive power of MF that have emanated from the research on recommender systems. One of these techniques is SVD++ [10]. The SVD++ tries to mix strengths of the latent model as well as that of the neighborhood model. It introduces a temporal dimension to the latent model on the premise that the accuracy of the model can be improved by taking into consideration changes in user preference as a function of time. We believe that such a model is more likely to improve the accuracy in the prediction of student marks. Another intervention that may be explored is combining MF and a classical user-item collaborative filtering technique. This was shown to improve the predictive power of MF in the work reported in [6].

REFERENCES

[1] A. Toscher and M. Jahrer, (2010). "Collaborative filtering applied to educational data mining," presented at ACM International Conference on Knowledge Discovery and Data Mining, KDD Cup Workshop. Available: ttp://www.commendo.at/UserFiles/commendo/File/KDDCup2010_ Toescher_Jahrer.pdf.

[2] A. A. Saa, (2015). Educational Data Mining & Students' Performance Prediction.,Interrnational Journal of Advanced Computer Science and Applications, 7(5), 2016., Available: https://thesai.org/Downloads /Volume7No5/Paper_31-Educational_Data_Mining_Students_Performance_Prediction.pdf

[3] Yehuda Koren , Robert Bell , Chris Volinsky, Matrix Factorization Techniques for Recommender Systems, Computer, v.42 n.8, p.30-37, August 2009 [doi=10.1109/MC.2009.263]

[4] Gower, S.: Netflix Prize and SVD pp. 1–10 (2014)

[5] Kirk Baker, "Singular Value Decomposition Tutorial", Unpublished.

[6] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," in Proc. The 1st Workshop on Recommender Systems for Technology Enhanced Learning, Elsevier's Procedia CS, 2010, pp. 2811-2819.

[7] Hwang, C.S., Su, Y.C.: Unified clustering locality preserving matrix factorization for student performance prediction. IAENG Int. J. Comput. Sci. 42(3), 245 (2015)

[8] A. Polyzou and G. Karypis, Grade prediction with models specific to students and courses, International Journal of Data Science and Analytics (2016)

[9] Ientilucci, E.J., (2003). "Using the Singular Value Decomposition". http://www.cis.rit.edu/~ejipci /Reports/svd.pdf

[10] Gower, S.: Netflix Prize and SVD pp. 1–10 (2014)

[11] Shahiri A.M., Husain W., and Rashid N. A. A Review on Predicting Student's Performance Using Data Mining Techniques. Proceedings of the Third Information Systems International Conference (ISICO2015). 2015.

[12] Gatsheni, N.G. and Katambwa, O.N., The Design of a Predictive Model for the Academic Performance of Students at University based on Machine Learning. Proceedings of the International Conference in Artificial Intelligence (ICAI17). 2017.

[13] J.P. Vandamme , N. Meskens & J.F. Superby. Predicting Academic Performance by Data Mining Methods, Education Economics, 15:4, 405-419. 2007.