Educational Data Mining Rule based Recommender Systems

Ghadeer Mobasher[1], Ahmed Shawish[1,2] and Osman Ibrahim[1]
[1]*Faculty of Informatics and Computer Science, The British University in Egypt, Cairo, Egypt*
[2]*Ain Shams University, Cairo , Egypt*

Keywords: Educational Data Mining, Decision Trees, Reduced Error Pruning and Rule based Recommender System.

Abstract: Educational Data Mining (EDM) is an emerging multidisciplinary research area, in which data mining techniques are deployed to extract knowledge from educational information systems to help decision makers to improve the learning process and enhance the academic performance of the students. The available studies mainly focused on predicting the academic performance based on demographic and study related attributes. Most of the previous work adopted the decision trees as one of the most famous data mining techniques to predict rather than extracting real knowledge that reveals the reasons behind student's dropout. On the other hand, there were other studies in the psychological track to measure the mental health score based on the educational environment. This paper proposes a complete EDM framework in a form of a rule based recommender system that is not developed to analyze and predict the student's performance only, but also to exhibit the reasons behind it. The proposed framework analyzes the students' demographic data, study related and psychological characteristics to extract all possible knowledge from students, teachers and parents.Seeking the highest possible accuracy in academic performance prediction using a set of powerful data mining techniques. The framework succeeds to highlight the student's weak points and provide appropriate recommendations. The realistic case study that has been conducted on 200 students proves the outstanding performance of the proposed framework in comparison with the existing ones.

1 INTRODUCTION

Educational Data Mining (EDM)helps decision makers to improve the learning process and enhance the academic performance of students in different educational programs by applying powerful data mining techniques . The greatest challenge in EDM is to identify the main factors affecting the student's academic performance and hence being able to better predict the student's achievement level in a pro-active manner (Prabha and Shanavas, 2014).

Most of the available EDM researches focused on two categories of data sets as inputs for analysis. Firstly, the data set related to the student's educational related attributes, such as the study and homework hours , Reading and Writing skills,etc.Second the demographic data set like gender, family related information, life style, etc. (Kovacic, 2010).

Despite the enormous researches done based on these two types of data sets, they only focused on exploring as much as possible attributes for the purpose of enhancing the accuracy of their prediction. Their results were mainly limited to predict the student's performance rather than reveling the real reasons behind it (Nasiri and Minaei, 2012).

It is also worth to note that most of the data sets of these researches were mainly acquired from one source only, which is the student's registration forms without any involvement of other parties like teachers and parents that participate in the student's learning process.

More and above, none of the previous studies included the student's mental health as part of the analysis. According to the Canadian Mental Health Association, approximately one in five children and youth has a mental health challenge that directly affects the student's learning capabilities. As a result, early identification is so critical and can lead to improvements in school and better health outcomes in life (Ramirez, 2014). Regarding the previous educational psychology researches that focused on assessing the student's mental health score based on the educational settings using statistical analysis, authors have analyzed the mental health based on the student's educational settings, to come out with results that reveal interesting directly proportional relationships between the student's educational settings of their study program and their mental health(Li et al., 2008).

Mobasher, G., Shawish, A. and Ibrahim, O.
Educational Data Mining Rule based Recommender Systems.
DOI: 10.5220/0006290902920299
In *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017) - Volume 1*, pages 292-299
ISBN: 978-989-758-239-4

We can easily notice the isolation between the two fields despite their correlation. It's crucial to add value to the identification of factors that have great impact on the student's academic performance with the usage of the powerful techniques of data mining to early alert weak students. Moreover, to provide a proactive decision making based on the extracted patterns to enhance the student's learning outcomes.

In this paper we propose a complete framework in the form of a rule based recommender system, which is developed not only to analyze the student's performance, but also to reveal the reasons behind it.In the proposed framework, we diversify the input data by combining the student's demographic data, educational activities and psychological characteristics. Moreover, the data set is not only extracted from the student's registration forms, but also from the teachers and parents to acquire their evaluation based on student's behavior at class and home.

The proposed recommender system will provide a proactive approach to early alert the issues that might affect the students academic performance in his/ her educational life. Moreover, it will provide the student, parents and the school with the best recommendation, that points out the weak points of the student that need to be considered with the appropriate best treatments based on the corresponding cases. This paper highlights three main contributions as follows:

- Concluding the most relevant attributes from the previous researches that have the highest influence on the student's academic performance.
- Deploy a set of very powerful data mining techniques to accurately predict the student's academic performance and extract all possible knowledge from the input data.
- Develop a recommender system to provide the student, parents and teachers with the appropriate recommendation based on the corresponding problems.

An extensive simulation studies have been conducted on a realistic real cases including 200 Elementary Students, revealing the outstanding performance of the proposed framework in comparison with the previous ones.

This paper is organized as follows: Section 2 illustrates the background. Section 3 introduces the proposed techniques and methodologies to develop the EDM Rule Based Recommender System. Section 4 describes the experiment and the results. Finally, Section 5 discusses the conclusion and the future work.

2 RELATED WORK

In this section we provide a comprehensive study of all the previous EDM Researches and we also review the available educational recommender systems.

2.1 Previous Predictive Models

The previous predictive models only focused on using the student's demographic data like gender, age, family status, family income and qualifications. In addition to the study related attributes including the homework and study hours as well as the previous achievements and grades. Bhardwaj and pal predicted the student's performance at the end of the semester through student's data such as attendance, assignments marks and class test marks (Baradwaj and Pal, 2012). From diverse literature, the observed predicted poor performance has been mostly traced to poor previous scores, demographic data as well as the level of intelligence(Ahmed and Elaraby, 2014). In addition to other background factors such as socioeconomic status, family education and occupation, religion and even ethnicity have been identified as factors that have highest implications on the student's academic performance(Berger and Archer, 2016). These previous work were only limited to provide the prediction of the academic success or failure, without illustrating the reasons of this prediction. Most of the previous researches have focused to gather more than 40 attributes in their data set to predict the student's academic performance. These attributes were from the same type of data category whether demographic , study related attributes or both , that lead to lack of diversity of predicting rules. As a result, these generated rules did not fully extract the knowledge for the reasons behind the student's dropout.

Apart from the previously mentioned work, there were previous statistical analysis models from the perspective of educational psychology that conducted a couple of studies to examine the correlation between the mental health and the academic performance. Previous researches studied the academic performance of students with respect to the correlation with their mental health behavior. Their results demonstrate that feelings of anxiety, depression, and time pressure negatively affected the performance of these students. However, participating in extracurricular activities alongside having a good support system positively affected the academic performance(McLeod et al., 2012). Moreover, The Center for Addiction and Mental Health found through surveys that final year students were least likely to report these symptoms as compared to students in other years (Weare and Nind,

2011).

After surveying the previous related work, we conclude that the previous attempt's main goal was the identification of the highest impact factors on academic performance. Some of them were from the educational data mining perspective , while others from the educational psychology methodology . Both fields have the same goal, but each of them have different methodology .

2.2 Previous Educational Data Mining Recommender Systems

Concerning the previous recommender systems, most of them are student oriented and they predict the performance of a specific subject or provide a single learning resource rather than predicting student's performance in each course and providing overall prediction to the student's final score. For example, Piedade (Piedade and Santos, 2008) that proposed a SRM model provide support in the form of an effective student institution relationship through monitoring of the students and their academic activities. Also,Kanokwan (KuyoroShade et al., 2013) built on the previously mentioned SRM model to propose an intelligent recommendation system framework for student relationship management that can assess the performance of the students and provide the appropriate recommendation for the choice of single course.

The existing systems mostly focused on predicting the student's performance in a particular course with respect to the student's data from the admission . Student's oriented systems have the purpose of aiding students to the best decision for specific course registration enrollment. The type of the recommendations was too brief, they missed illustrating the methodologies to apply them.

We believe that the real recommender system should provide the suitable advises to enhance the student's overall academic performance based on the type of the challenges that the student is suffering from during the academic year.

3 PROPOSED FRAMEWORK

This section describes the proposed framework through four sub sections. First, the new combination of input data set is explained. As well as, detailed illustration of data pre-processing is presented. Second subsection describes the data analysis and processing. Finally, the implementation details of the proposed framework is discussed.

3.1 Input Data Model

The proposed framework firstly focuses on merging the demographic and study related attributes with the educational psychology fields, by adding the student's psychological characteristics to the previously used data set (i.e., the students' demographic data and study related ones). After Surveying the previously used factors for predicting the student's academic performance, we picked the most relevant attributes based on their rationale and correlation with the academic performance (Osadan and Burrage, 2013) (Cimmiyotti, 2013). Concerning the student's psychological characteristics, they were extracted through group of symptoms that asses the student's mental health based on children fact sheet from National Alliance on Mental Health (NAMI) (Deb et al., 2015). According to NAMI, the top 5 mental illness that the school student's might suffer from are Anxiety, Hyperactivity, Bipolar, Conduct, and Reactive Attachment Disorders. The previous studies proved that low mental health score has negative academic performance implications

To remove any unnecessary attribute, we conduct a feature selection analysis using Information Gain as an attribute evaluator and Ranker as a search method. The feature selection analysis dropped 8 attributes from the data set, that includes 45 attributes as they report zero gain information gain.

Three questionnaires have been designed to collect the above listed data for the student, parents and teachers. The first one was questioned to the student, mainly focusing on the set of pre-mentioned demographic data. The other two questionnaires, were targeting parents and teacher's evaluation for the student's psychological characteristics and study related skills.

The input model is divided into three main categories, the first is the demographic data which is illustrated in Table 1. The second category is the study related attributes illustrated in Table 2. Finally, the third category is psychological characteristics shown in Table 3.

3.2 Data Analysis

This paper proposes a hybrid system which is based on data mining techniques using classification algorithms, specially focusing on four rule induction algorithms One R, Zero R, JRIP and PART. And four decision algorithms J48, Random tree, REP tree and Decision stump. The first four classifiers belong to Rules and the second ones belong to Decision Trees. These classification algorithms belong to the "white

Table 1: Demographic Arributes.

Attribute Name	Expected Values
Stage	1,2,3,4,5,6 Primary
Gender	Male, female
Home Type	Flat, villa
Father qualification Mother qualification	School level, BSc. MSc PHD
Family income	Poor, Middle, High
Parental status	Married, Divorced, Separated.
No. of siblings	Numeric

Table 2: Study Related Attributes.

Variable Name	Variable format
Number of courses	Numeric
Homework hours	
Study hours	
Counting skills	Poor, Average, Excellent
Arithmetic knowledge	
Reading skills	
Writing skills	
Hand-writing skills	

box" classification model and can be used directly for decision making. To validate that our proposed work outperform the previous related work, after merging between the demographic, study related and psychological characteristics and after applying 8 different powerful data mining techniques to predict the academic performance.

In order to choose the best data mining technique, there are three main concerns for the model evaluation : the tree size, tree complexity and prediction accuracy of the decision tree.

The proposal aims to analyze student's demographic data, study related details and psychological characteristics in terms of final state to figure whether the student is on the right track or struggling or even failing. In addition to extensive comparison of our proposed model with the other previous related models.

3.3 Framework Implementation

The proposed recommender system is implemented with java language using Net Beans version 7.3 as a java environment. WEKA Java API is be used to implement the data mining and machine learning algorithm. The intended user is the Data Analyst. The recommender system enables the data analyst to upload the data set and then view the predicted set of

Table 3: Psychological Attributes.

Variable Name	Variable format
Talkative	Yes/No
Lack of motivation	
lack of organization	
History of frequent suspension	
Mood swinging	
Frequent Abscence	
Failure to finish work	
Bossy	
Easily distracted disability to listen	Always, Never, Often
Blurt Out	
Trouble playing quietly	
Makes a lot of mistakes	
Lying	
Physical fights	
Stealing	
Unusual speech patterns	
Toileting issues	
Refusal to answer simple questions	
Memory abilities	Poor, Average, Excellent.
Attention abilities	
Eye-contact abilities	

rules and the generated tree. Moreover, the data analyst can add student's information to predict the overall academic performance, the predicted value will be displayed. The recommender system isn't only limited to the prediction of state, but also provides the student, parents and teachers with the corresponding recommendations and strategies to follow to improve the learning outcomes. These recommendations are based on experimented studies for enhancing the student's academic performance. In addition to the mentioned above functionalities, the System will also alert all parties with the possible upcoming mental illnesses that the student might suffer from. Due to the gathering of the psychological characteristics that formulate set of symptoms to couple of mental illnesses that might negatively impact the academic performance, as mentioned in section 2.

4 CASE STUDY

This section presents the realistic case study and discusses its results. The Student's academic performance prediction and recommendations will be illustrated with a comparative analysis between our proposed framework and the previous related work.

4.1 Experiment Settings

Once the questionnaires have been acquired from the pre-mentioned sources and after the data set passed through the pre-processing phase. The experiment is conducted on 200 elementary students , after eliminating the inconsistencies between Teachers and Parent's Response on the study related data and psychological characteristics.

The data is then analyzed for knowledge extraction; the data is first classified on the hybrid 8 classifier's algorithms using 10- fold cross validation method. Data were randomly divided into 10 parts. Each held holds the next and learning scheme trained nine-tenth of the rest of the data set, then the error rate is calculated in the holdout set. The learning procedure is performed 10 times on different random training set. Finally, an average of 10 error estimates to produce estimates of the overall error. Detailed result analysis will be discussed in the coming subsection.

4.2 Predictive Model Results and Validation

In terms of prediction accuracy, REP Tree outperform with a prediction accuracy of 73.6 %. While the least prediction was for Zero R with 45.5 % as prediction accuracy. Detailed bar chart graph is illustrated in figure 1.

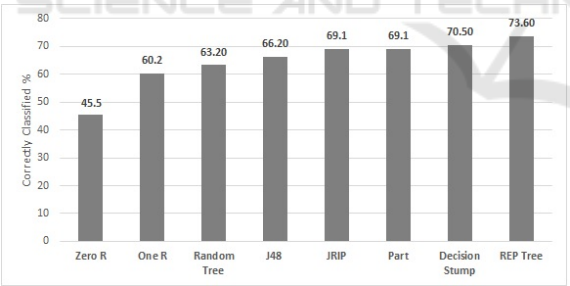

Figure 1: Classifier's Prediction Accuracy.

Regarding the classifier's tree size, REP tree has the smallest tree size. While, Random Tree has the largest tree size. The complexity of the tree is measured by the size of the tree, the total number of leaves and the tree's depth. The less the tree size, total leaves and tree depth the less complicated the tree. In figure 2, detailed comparison between classifier's tree size and complexity.

Although Zero R has the least tree size and complexity, however it is the least reliable classifier to apply, because the extracted rules lost structural complexity. Only one rule is generated to predict the academic performance.

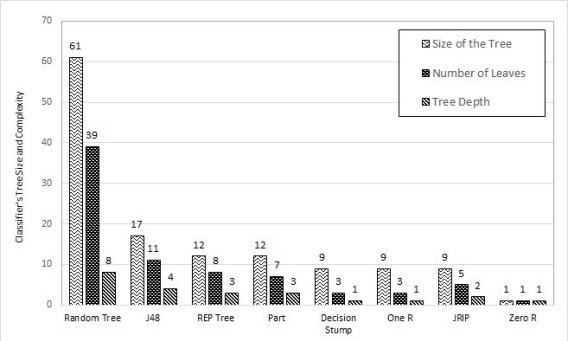

Figure 2: Classifier's tree size and complexity.

It is very clear to conclude that REP tree is the best classifier to apply.As the size of the REP tree is the smallest with the lowest complexity and highest prediction accuracy. REP tree is a post-pruning method; this decision tree finds the smallest version of the most accurate sub-tree with respect to the pruning set. During the pruning the size of the tree will get reduced and reduces the complexity and thus improves the prediction accuracy. Regarding the Rep's prediction accuracy results analysis 73.5 % were predicted true when they were actually true, while 12.5 % were predicted true however, they were actually false. In Table 4 , the generated REP Decision Tree.

Table 4: Generated Rep Rules.

Generated Rules	Predicted State
If the Arithmetic Knowledge == Excellent	Then the state == On the Right Track
If the Arithmetic Knowledge == Average and Reading Skills==Excellent == Excellent	Then the state == Strugling
If the Arithmetic Knowledge == Average and Reading Skills==Poor	Then the state == Failing
If the Arithmetic Knowledge == Average and Reading Skills== Average and History of Frequent Suspension==Yes	Then the state == Failing
If the Arithmetic Knowledge == Average and Reading Skills== Average and History of Frequent Suspension==No	Then the state == Strugling
If the Arithmetic Knowledge == Poor and Makes lots of Mistakes ==Always	Then the state == Failing
If the Arithmetic Knowledge == Poor and Makes lots of Mistakes ==Never	Then the state == Strugling
If the Arithmetic Knowledge == Poor and Makes lots of Mistakes ==Often	Then the state == Strugling

If we applied the previous related research strategy that used the demographic data to predict the academic performance on our data set. The best classifier is J48 with a correctly classified instances and the predicted rules mainly depend on the fathers and mother's qualifications, home type, family income and frequent absence. While if we used the demographic data and the study related attributes to predict the academic performance, the best classifier is the REP tree. Arithmetic knowledge, Number of courses, Writing Skills, Homework Hours and stage have the highest impact on the prediction of performance.

Finally, if we applied the third strategy of using the psychological characteristics to predict the performance, the best classifier was J48 correctly classified instances. History of frequent suspension, Failure to finish work, blurt out, Memory Abilities, Lying and disability to listen were mainly considered in the prediction of performance.

In figure 3, detailed comparative prediction accuracy between old strategies vs our work.

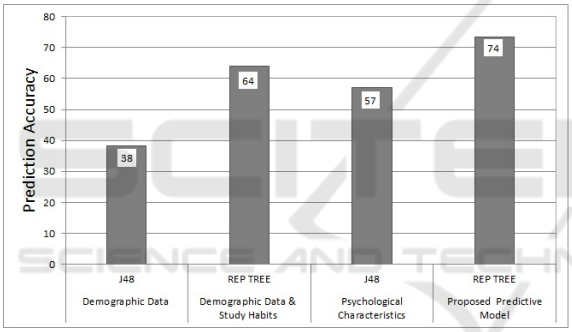

Figure 3: Comparison between previous work and our model's prediction accuracy.

Beyond the out performance of our proposed strategy in terms of accuracy percentage and our extracted predictive rules , which are consistent with the theoretical aspects of the educational principles. We can assert that our work added great values to the prediction of the academic performance. After applying our proposed methodology to merge between educational data mining and educational psychology perspectives, new results have been generated that better predicted the student's academic performance.

4.3 Recommeder System Discussion

The data analyst first logs in and browses for the file to upload it. Then the system automatically applies the 8 classifiers and uses the highest best classifier's prediction accuracy, which is REP Tree in our case. Then the data analyst can have the option to see the classifier's prediction rules and generated decision tree as well. Also a detailed message is displayed with detailed classification's accuracy. In addition to the functionalities mentioned above, the data analyst can enter the data of the student who wants to predict his/her overall academic performance. After the analyst fills in the student's information, the system displays the predicted academic state of the student and presents the appropriate recommendations for the student, the parents and the teacher.

Moreover, it alerts all parties with the mental disorder that the student might face in the future. It is possible that the student might not face any academic problems currently, but in the future, the student will be facing academic performance challenges due the mental illness disorder. So the system will not only provide the analyst with the predicted academic state and the appropriate recommendations, but highlights any upcoming mental illness that the student might shown in the future. Snippets of the system's interface are illustrated in figure 4, which presents an example of the parent's , teacher's and student's recommendation after entering student's data and predicting performance. The proposed system avoided the

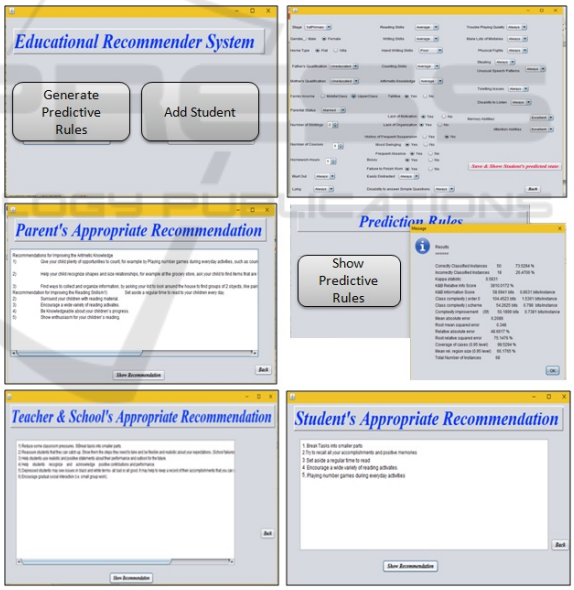

Figure 4: Screenshot of framework interface .

drawbacks of other system. It provides recommendation for all parties that participate in the student's educational life, including the student, the teachers and parents. Samples of the recommendations addressed to the student are illustrated in table 5. Table 6 illustrates sample of recommendations for the teachers and parents.

The extracted recommendations are based on proven classroom fact sheets (for Disease Control et al., 2013) (Hite, 2009) (De Corte et al.,).

Table 5: Samples of the suggested recommendations for Students.

Problem	Suggested Recommendation
The student's arithmetic knowledge and reading skills are average or poor	• Play arithmetic games so that you can have fun and at the same time learn and practice. • Write the small steps as a draft while solving rather than memorizing the numbers to avoid mistakes. • Scan before you read. • Practice, the more you read the better reader you will become and smarter too, so feed your mind.
The student is suffering from AD/HD ,(s)he is easily distracted ,blurt out ,poor attention and memory abilities,talkative ,lacks organization , having trouble to play quietly and makes lots of mistakes	• Enhance your social skills and try to get involved with other peers. • When you feel low or sad, try to divide your studies into small manageable parts.

Table 6: Samples of the suggested recommendations for Parents and Teacher.

Problem	Suggested Recommendation
The student's arithmetic knowledge and reading skills are average or poor	• Implement a coherent reading program level. • Focus on fluency and comprehension. • Create a culture that encourages learning, thinking, reflection and self-analysis
The student is suffering from AD/HD ,(s)he is easily distracted ,blurt out ,poor attention and memory abilities,talkative ,lacks organization , having trouble to play quietly and makes lots of mistakes	• Provide the student with recorded books as an alternative to self-reading when the student's concentration is low. • Break assigned reading into manageable segments and monitor the student's progress, checking comprehension periodically. • Devise a flexible curriculum that accommodates the sometimes rapid changes in the student's ability to perform consistently in school. • When energy is low, reduce academic demands; when energy is high, increase opportunities for achievement. • Identify a place where the student can go for privacy until he/she regains self-control.

Moreover, most of the previous attempts were focusing on predicting the student's performance in a specific course with respect to the student's information gathered from the admission. It only focuses on identifying the factors that have high impact on the student's success or failure in certain subject. Based on the no-free theorem, the system that was able to predict students performance in a particular course may not do so in the overall performance of the students at the end of each academic year.

One of the previous recommender system attempts, let the user add the IF-ELSE rules manually and this of course increases the human error. while the implemented system let the data analyst loads the data set, enabling dynamic prediction rules extraction, even if the data set changes.

5 CONCLUSION AND FUTURE WORK

This paper proposes a complete EDM framework in the form of a rule based recommender system that analyze the student's academic performance to point out the student's weak points and provide appropriate recommendations for treatment.The paper combined three major factors that affect the student's academic performance: demographic data, educational related attributes and psychological characteristics. The realistic case-study conducted on 200 students assured the outstanding capabilities of the academic performance prediction, depth of knowledge extraction, and great benefit of recommendations provided by the proposed framework in comparison with the available work. The generated rules also showed a perfect matching with the scientific proved facts. The diversification of the data sources and the involvement of teachers and parents present a notable improvement in this work.

Regarding the future work, a larger data set needed to be used in different academic stages. Moreover, new psychological characteristics need to be added with the supervision of professional psychologist to better discover new patterns and enhance the prediction results.

REFERENCES

Ahmed, A. B. E. D. and Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, 2(2):43–47.

Baradwaj, B. K. and Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

Berger, N. and Archer, J. (2016). School socio-economic status and student socio-academic achievement goals in upper secondary contexts. *Social Psychology of Education*, 19(1):175–194.

Cimmiyotti, C. B. (2013). Impact of reading ability on academic performance at the primary level.

De Corte, E., Walberg, H., Fraser, B., Kirst, M., Teichler, U., and Wang, M. The international academy of education.

Deb, S., Strodl, E., and Sun, J. (2015). Academic stress, parental pressure, anxiety and mental health among indian high school students. *International Journal of Psychology and Behavioral Sciences*, 5(1):26–34.

for Disease Control, C. et al. (2013). Make a difference at your school.

Hite, S. (2009). Improving problem solving by improving reading skills.

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

KuyoroShade, O., Oludele, A., Okolie Samuel, O., and Nicolae, G. (2013). Framework of recommendation system for tertiary. *Framework*, 2(04).

Li, H., Li, W., Liu, Q., Zhao, A., Prevatt, F., and Yang, J. (2008). Variables predicting the mental health status of chinese college students. *Asian Journal of Psychiatry*, 1(2):37–41.

McLeod, J. D., Uemura, R., and Rohrman, S. (2012). Adolescent mental health, behavior problems, and academic achievement. *Journal of health and social behavior*, page 0022146512462888.

Nasiri, M. and Minaei, B. (2012). Predicting gpa and academic dismissal in lms using educational data mining: A case mining. In *6th National and 3rd International conference of e-Learning and e-Teaching*, pages 53–58. IEEE.

Osadan, R. and Burrage, I. A. (2013). The effect of age, societal status and sexuality on students elementary schooling.

Piedade, M. B. and Santos, M. Y. (2008). Student relationship management: Concept, practice and technological support. In *2008 IEEE International Engineering Management Conference*, pages 1–5. IEEE.

Prabha, S. L. and Shanavas, D. A. M. (2014). Educational data mining applications. *Operations Research and Applications: An International Journal (ORAJ)*, 1(1).

Ramirez, J. (2014). The relationship between school-based mental health services and academic achievement.

Weare, K. and Nind, M. (2011). Mental health promotion and problem prevention in schools: what does the evidence say? *Health promotion international*, 26(suppl 1):i29–i69.