Performance Evaluation and Comparison of Classification Algorithms for Students at Qassim University

Ayshah Abdulrahman Al-Noshan¹, Mohammed Abdullah Al-Hagery², Hessah Abdulaziz Al-Hodathi³, Meznah Sulaiman Al-Quraishi⁴

¹, ³Master Student, Department of Computer Science, Collage of Computer, Qassim University, KSA
²Associate Professor, Department of Computer Science, Collage of Computer, Qassim University, KSA

Abstract: The level of the student may change at the undergraduate level from his level in intermediate or secondary school. Educational and personal factors may affect the performance of university students, especially in the first year of university. In this study, we measure a relationship between a Grade Point Average (GPA) of the students at Qassim University in the first year, and variety of personal and educational factors of the students, such as parents education, number of hours using smartphones and the degree of Scholastic Achievement Admission Test. To get the results, we used six different classification algorithms to measure accurately for attributes between these algorithms and find the best algorithm that gives us a high accuracy percentage for prediction of student’s performance. We apply these algorithms using RapidMiner platform, which is one of the most famous and strong data mining tools. Based on the overall comparison of these six classification algorithms, the Neural Network algorithm had the highest accuracy percentage 84% and the lowest was 56.55% of the AdaBoot algorithm. Then, we apply association rule to check if some rules demonstrate the existence of direct relationships between personal and educational factors and student performance or not.

Keywords: Educational Data Mining, Student’s Performance, Classification Algorithms, Association Rule, Personal Factors

1. Introduction

Nowadays, educational institutions and universities seek to provide a suitable environment for improving student performance. Prediction is one way to predict the factors that affect a student’s educational level and extract knowledge from a dataset by using data mining techniques. The application of data mining techniques in education is known as Educational Data Mining (EDM), which is an important process that enables us to extract the data patterns to discover hidden knowledge and key characteristics from students’ datasets. The EDM has a wide interest in recent years from researchers and universities and is encouraged to improve the performance of their expected level and focus on the factors affecting students and help students themselves to enhance efficiency and achievements. Also, it may help universities to reduce the education process expenses [1].

Qassim University is one of the universities in Saudi Arabia that seeks to provide a suitable learning environment for its students and enhancing the learning process. Moreover, it is interested in the guide the students’ learning and improving the performance for them. The university has about twelve colleges in different places and includes 71 scientific majors in different fields. The university offers all degrees after colleges in different places and includes 71 scientific majors in different fields. The university has about twelve colleges in different places and includes 71 scientific majors in different fields. The university offers all degrees after

2. Related Work

Bakhshinategh et al. in [3] reviewed some studies different tasks and applications that use educational data mining and classify them into different categories based on their ultimate goal of the application, similarity, and differences among them. While one of these categories was student modeling which we will focus on in our paper.

In [4], Mueen et al. was using Naïve Bayes, Neural Network, and Decision Tree classification algorithms to modelling and predicting students’ academic performance. The authors used a different variety of information varies from general information of students. The final source of
information was academic information, which depends on the Grade Point Average (GPA) and other factors. While Devasia et al. [5] showed that study depends on predicting the students’ performance by using the Naïve Bayes model. The study information depends on the educational factors such as time of study and hours spends on the study and non-educational factors which effects on the students’ performance such as mother's education and using of the social network.

Ramaswami et al. in [6] used four different classification algorithms and process mining features to increase the prediction accuracy of student academic performance. These algorithms are Naïve Bayes, Random Forest, Logistic Regression and K-Nearest Neighbours. The researchers collected data comprised of activities from an educational tool called Xorro-Q. The result of their experience was that Random Forest had the highest accuracy.

Sarra et al. worked in [7], to evaluate the usefulness of the Bayesian profile regression to identify and detect the main characteristics of the students drop out of their study. They collected information about an undergraduate student by a questionnaire filled. This study of this paper focused on two factors that may affect the dropout: internal factors, which are related to personality, and external factors, which are related to the other factors such as educational.

Oyedotun et al. in [8] used the Neural Network to predict the number of times that the student will have to repeat the course and to know what is the factors that most effect on the students’ success. Also, Burgos et al. in [9] proposed knowledge discovery techniques to analyze historical student grades that depend on different factors such as week number. Besides, Hussain et al. in [10] evaluated the performance of university students in India used different classification algorithms to prevent the student from the dropout. However, Livieris et al. in [1] used classification algorithms such as MLP to provide a user-friendly software that accurately identifies students who are at-risk of failing in the final exams during the academic year and alert them to improve their performance.

Ahmed and Elaraby in [11] used the predicting for the performance of students using decision tree (ID3) classification method to predict the final grade of student's performance. However, Yadav et al. in [12] used different Decision tree algorithms to predict the final performance at the end of the course while the student information was used in this study is like assignment marks. On the other hand, Hajizadeh et al. in [13] used data mining techniques to evaluate the performance of faculty members to improve the effectiveness of their performance at the end of each semester.

While the prediction of student's performance in [14] performed by Sorour et al. was by making mining for the comment that the students wrote after every lesson introduced. In Error! Reference source not found., Delfino studied the performance of students based on student engagement and check if there is a relation between them or no. He used multiple things to treat and analyze data such as mean and ranking and multiple regressions. Also, Razaque et al.[15], predicted the academic performance of bachelor academic students in Pakistan. They applied two data mining techniques are classification and clustering to discover knowledge. The study observed a relationship between student’s attendance lecture & discussion and academic performance.

Almarabeh in [17] used a student college database to predict students’ performance by using several classification techniques such as Naive Bayes and Neural Network. However, discovering the subgroup was used by Halal et al. [18] to study the essential factors that affect the academic performance of students. While Fernandes et al. in [19], based on the Cross-Industry Standard Process for the data mining model to predict the performance of students in public schools in Brazil. Also, the performing of predictive analysis by Amrieh et al. in [20] was using ensemble methods to improve the performance of classifiers while they based on features of student's behavioral. In addition, the study of Gopaldas in [21] aimed to find variables affecting academic performance. They used correlation technique and regression. The result of this study showed that gender, parent income and attendance in class are the most factors affecting academic performance.

In paper [22], Widyaningsih et al. used a semi-supervised learning technique to classify and cluster the mathematical student performance at the University of Indonesia. The study depends on 140 samples collected by questionnaire from the first-year student with 27 features applied. In this paper, the author used K-mean clustering and Naive Bayes classifier to show the result. This study showed that the K-mean clustering and Naive Bayes classifier algorithm could be used to classifying the performance of the student.

A studying the acceptance for EDM for undergraduate students was the focus for Wook et al. in [23]. They focused on it to increase achievements, to predict performance and to measure some of the outcomes. Also, in [24], Asif et al. analyze the performance of undergraduate students using several classifiers and study the progress of students using clustering and then combine these progresses with the prediction results.

In [25], Weldegebriel et al. in their study evaluated the academic performance by examining different determinants for students of grade 10 but specifically in Tigray by using the estimation of Multiple Linear Regressions and Ordered Probit models. They conclude that students vary in their academic performance based on their sex, residence and other factors.

Ab Razak et al. in [26] used different ways of analysis to find if there is a relationship and to identify essential factors that may affect the academic performance in a Higher Learning Institution. They collected data using questionnaires manner and analyzed by Statistical Package for the Social Science (SPSS). They presented different analyzed such as multiple regressions and reliability analysis. While Anazia in [27], focused on their study to predict the performance of students in Senior Secondary School in
Economics based on Quantitative and Verbal Aptitudes. They used ANOVA and multiple regressions of SPSS to analyze the data. They found that quantitative aptitude might affect more on the performance of students than verbal aptitude. Bonsaksen and Ellingham in [28] identified factors affecting on the second-year undergraduate occupational therapy students in Norway used SPSS to analyses the statistical performed. The study was collected by questionnaires where it contains two sections of question are sociodemographic backgrounds like age and education-related variables. The authors have found that there is a relationship between a high age and better academic performance. Moreover, the study in paper [29] by Adeyemi et al. searched about the influence of peer group students in academic performance at the selected department of Babcock University. A generated data was analyzed by using SPSS. The result of this study showed that the student affecting the performance of undergraduate students.

Hussain et al. in [30] presented a framework to make accreditation a more efficient and effective by apply educational mining such as prediction.

The study by Khan and Ghosh [31] used association-mining approach to find the association between the quality of teaching and the improvement of student's performance.

On the other hand, Fabito et al. in [32] employed a quantitative approach to examine the students’ academic performance associated with cyberbullying and mobile game addiction. Besides, the data were collected from Google's questionnaire, which contained 21 questions about gaming addiction. Where eighty-five respondents were from the Philippines. The results of this study were neglect mobile game addiction as a factor for academic performance. Also, Alalwan et al. in [33] provided three theories to develop a model of factors that affect student performance in higher education. The authors aimed to show the relation of using social media for collaborative and learning and student performance. This study applied on 863 undergraduates’ students by using constructivism theory, TAM Model and communication theory. The result showed that the use of social media for collaborative and learning affecting student performance.

3. Tools and Techniques

In this study, we use RapidMiner as a tool to deploy different data mining models on it. The techniques that are used in this study are six different classification algorithms that are Neural Network, K-Nearest Neighbours, Decision Tree, AdaBoost and Multi-Layer Perceptron. In addition, we use an association rule technique.

4. Methodology

The analysis in this study is applied to show the relationship between the personal and educational factors of students at Qassim University, and the performance of students in the first year based on the classes of GPA in Qassim University using classification algorithms for prediction and association rules. A data-mining model was deployed on RapidMiner platform using six different classification algorithms: Neural Network, K-Nearest Neighbours, Decision Tree, Random Forest, AdaBoost and Multi-Layer Perceptron algorithm. Moreover, the association rule technique is applied after we measure the accuracy of prediction to check if some rules exist that demonstrate the existence of direct relationships between personal and educational factors and student performance or no based-on GPA of student. The analysis of the dataset in this study was for 724 different students who got it from the questionnaire. As we got our dataset from a questionnaire that was by using google forms, we select the sample of our dataset using linear sampling for records that ordered randomly by taking 80% of data to be for train and the remainder 20% was for test the performance of the model.

5. Description of The Dataset Analyzed

In this section, we present a brief description of the dataset analyzed in this study that is focused on Qassim University students. The attributes of the dataset analyzed in this study combine some of the educational and personal factors that may affect the performance of students in the first year, and GPA attribute that considered as label for prediction. Educational factors that we are focused on it are Intermediate score, Secondary score, Scholastic Achievement Admission Test (SAAT) score, General Aptitude Test (GAT) score and type of studying if it is scientific or not. On the other hand, the personal factors of students such as gender type, number of hours spend on smartphones, type of high school if it is governmental or not, education of parents, the degree of English language for the student and number of student's desire accepted at university. The GPA attribute that used as label for prediction for student grades in Qassim University fall into the following categories, 4.50 - 5.00, 3.75 - 4.49, 3.74 - 2.75 and 2 - 2.74.

Figure 1 (a), (b), (c), (d), (e) and (f) for visualization of data are illustrated the relation between these features and GPA. The colors shown in Fig. 1 indicate the categories of GPA score such that light blue color is the class of GPA score whose degree is 4.5 – 5. The green color is the class of GPA score whose degree is 3.75 - 4.49, orange color is the class of GPA score whose degree is 2.75 - 3.74 and black color is the class of GPA score whose degree is 2 - 2.74.
6. Results and Discussion

When we apply prediction for the performance of the students in Qassim University in the first year, we got some effective results. The accuracy for six different classification algorithms for the prediction of students' performance was varying from one to another with all students' features deployed. The highest accuracy was 84.83% that is using the Neural Network algorithm. Then, the order for the accuracy of the remaining algorithms was the Random Forest algorithm with an accuracy of 83.45%. The third best performance was Decision Tree with an accuracy of 81.38%. The fourth best performance was Multi-Layer Perceptron with an accuracy of 80%. K-Nearest Neighbours was the fifth-best performance with the accuracy of 73.79% and the last best performance was AdaBoot with the accuracy of 56.55%. The performance comparison for six different classification algorithms is illustrated in Table 1.

<table>
<thead>
<tr>
<th>Algorithm/measure</th>
<th>Accuracy Percentage</th>
<th>Error Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>83.45%</td>
<td>16.55%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>81.38%</td>
<td>18.62%</td>
</tr>
<tr>
<td>K-Nearest Neighbours</td>
<td>73.79%</td>
<td>26.03%</td>
</tr>
<tr>
<td>Multi-Layer Perceptron</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>84.83%</td>
<td>15.17%</td>
</tr>
<tr>
<td>AdaBoot</td>
<td>56.55%</td>
<td>43.45%</td>
</tr>
</tbody>
</table>

From the accuracy results shown above in Table 1 for different classification algorithms that predict the student’s performance, we observe that there is a good relationship between personal and educational factors and the GPA for Qassim University students. In addition, we observe the features that we are selected are affected by the student’s performance at Qassim University in the first year.

After we applied these algorithms, we apply association rule technique to check if there are some rules demonstrate the existence of relationships between personal and educational factors and student performance in the first year like if there is a relation between specific factors and GPA for the student in the first year or no. The results that we got it after applying association rule show that the performance of the student is affected by more and different factors and there are no specific factors that affect directly on the student’s performance more than the other. Unfortunately, the rules that we got it only demonstrate some relations between factors itself and not demonstrate the relation for these factors with GPA for students like these rules shown below:

1) English Degree for Student, From student graduate ---> Gender
2) Number of student's desire accepted at university ---> English Degree for Student, Gender
3) From student graduate ---> English Degree for Student, Gender

As we observed from working on RapidMiner environment to find the accuracy of prediction for the performance of the student in the first year in university based on educational and personal factors, we found that the highest accuracy was 84% and lowest accuracy was 56.55%. This accuracy indicates that there is a good relationship between educational and personal factors of students and the performance of the student in the first year in university.

However, the relationship between educational and personal factors and performance of the student in the first year was good but not high relation. Therefore, we observe this not high relation is because some of the educational or personal factors have a high relation and some do not have a high relation with the performance of students in the first year. For instance, the performance of the student in the first-year is increasing and highly affected by the increase in educational factors such as Intermediate score, Secondary score, Scholastic Achievement Admission Test (SAAT) score, and General Aptitude Test (GAT) score as shown in Figure 1 (a), (b), (c), and (d). In addition, the performance of the student in the first year is affected by personal factors such as the education of parents as illustrated in Figure 1 (e). Where Figure 1 (e) shows that most parents of students were educated. Where if both parents are educated, the performance of the student will be better because of the effect of a family environment. On the other hand, we observe some of the features don't affect significantly the performance of student such as the number of hours for using smartphones varies from one score to another as shown in Figure 1 (f). For example, some of the students use smartphones many hours and get a high GPA score that is 4.5 – 5, because it may divide their day properly or they use it for educational purposes to increase their performance like in research or for solving homework.

From the results, we observe that the features of students used in this paper focused on educational and personal factors, these factors maybe not enough because there are many other factors that may affect on the performance of student. For example, outcomes of the university, the presence of previous training courses, student's social status, teacher efficiency in intermediate and secondary school and so on. So, the selection for features to evaluate the prediction of student’s performance need to be improved and increased, because some of the features may affect the performance of the student and it is hard to get it from students. Some of these features are student skills, Teacher skills, degree of

![Figure 1: Dataset Analysis](image-url)
lecturers and family budget because many students do not know their family budget and if the student has a high relation with many friends or no.

We conclude from our work that the performance of the student in the first year in university is not limited to only the performance of the student in Intermediate, Secondary school, and personal factors. The performance is related for many things such as university policies, the efficiency of teaching staff, university activities, admission criteria for each university and other factors related to the student inside or outside the university or related to the university as a whole.

Finally, from the questionnaire that we did it, we asked students about the consequences that they faced in the first year to be as developed in the future. Most of the consequences that students faced were sudden pressure in the academic work that they did not face during studying in Intermediate or high school. Furthermore, they did not accept their desire for studying at university. Therefore, this encourages developing and improving the plan of studying in Secondary school and conducting development courses for pre-university students that will affect the performance of the student in university.

7. Conclusion and Future Work

Qassim University is one of the universities in Saudi Arabia that is interested in improving the performance of its students and improves performance measurements that is using Educational Data Mining. EDM is a science that interests in extracting knowledge from students’ dataset. This study measures certain attributes that may affect the education level of the student and to predict at-risk students who may have a lower level in their first year of study than their secondary school, to provide a special support to them. We measured eleven attributes of different aspects of a student’s life. In addition, we applied six classification algorithms and association rule by using RapidMiner tool to determine the relationship between these attributes as the educational and personal situation affects the level of students. We obtained that algorithm with Neural Network has the highest accuracy while AdaBoot algorithm has the lowest accuracy for the prediction of student’s performance. After that, we applied the association rule to discover interesting relationships between personal and educational factors and student performance. The result was the performance of the student is affected by more and different factors.

Finally, in future work, we will take a large sample of students from different universities in Saudi Arabia and apply it to a large dataset to get results that are more accurate. Where if we apply Neural Network that has the best accuracy on different universities, we will increase the performance of prediction because the performance of students may affect by the university policies, quality of staff and outcomes of each university. In addition to classification, we will use clustering to discover a hidden pattern that may be affected on student performance.

References


