Volume 15. Issue 2. Year 2021. Pages 91-100

Edukasi

https://journal.unnes.ac.id/nju/index.php/edukasi

Predicting Student Performance using Data Mining Techniques in Libyan High Schools

Mahjouba Ali Saleh[1*], Sellappan Palaniappan[1], Nasaraldeen Ali Alghazali Abdalla[2]

[1]University of Science and Technology Kuala Lumpur, Malaysia
[2]Bani Waleed University Libya

DOI: http://dx.doi.org/10.15294/edukasi.v15i2.30068

Info Articles

History Article
Submitted 2021-04-30
Revised 2021-07-26
Accepted 2021-11-14

Keywords:
Data Mining, Performance Predicting, Education, electronic systems, E-Learning

Abstract

Student performance in schools have been always the key factor for the teacher ability to teach and what brings good reputation to the school. Recently schools in Libya are facing an issue trying to figure out why students perform poorly in certain subjects and how can they know how they will perform next in the future in coming semesters in perspective subject. There are several methods proposed to predict the student's performance, using data mining. This paper proposes using Math and English as key factors to predict the performance of the students. results and findings of the presented method in terms of predicting students' performance based on their grades in Math and English. The results are divided in to three main sections clustering analysis using k-mean algorithm, classification analysis was done using two rounds first using Gain Ratio Evaluations to find out the top attributes that used by J84 algorithm in second round of classification, and rule association analysis using A priori algorithm. Rule association analysis is applied for the clusters generate by clustering analysis to generate the rules associated with each cluster. For each section, a list of inputs is presented with the scale used for the values followed by the results of the algorithm and explanation for the finding.

[*]Alamat Korespondensi:
E-mail: mahjouba.ali@phd.must.edu.my

p-ISSN 0852-0240
e-ISSN 2746-4016

INTRODUCTION

Student prediction performance is crucial. The prediction of student's performance has started when student's performance was not going up to standards. Therefore, schools had to find a way to fix the poor performance issue. One of the ways to fix this problem is implementing a system based on data mining that is depending on input from authorities, such as students background, previous performance in other schools and other subjects, the hours taken to study certain subject, marks of exams of subjects. Each methodology has its own weakness and strength in terms of predicting student's performance (Costa, Fonseca, Santana, de Araújo, & Rego, 2017). In this paper we are proposing students predicting performance based on their performance in exams and quizzes all along the semester and the academic year. Each student is taken individually to test the performance and determine the ability to predict future performance (Ahmed, Rizaner, & Ulusoy, 2016).

The core subjects taken prediction were English and Math. They are the key factors to predict the performance of Physics, Biology, Information technology, and Chemistry. The reason of taken Math and English as core subject to predict the performance is they participate in all subjects whereas physics and chemistry are based on numbers and English principles, also IT is all about English technology, so these two subjects have become the core factor for the test. This study focuses on predicting the student's performance based on their grades in Math and English. Data of student's exams and quizzes in all semesters will be taken for analysis to predict if the student will perform poorly of successfully in the future semesters and years. The subjects that are used for the comparison and prediction are Physics, Chemistry, Biology, and Information Technology (Ahmed et al., 2016). These subjects are cross compared against Math and English to determine the future performance of the student to prevent students from failing and changing school or major of study. Table1. Presents the elements used for this study.

METHODS

This research has proposed prediction using primary subjects that are Math and English. Math is used to predict the performance of the students, with using the quizzes and monthly reports. Math results were used for comparison with other subjects to estimate the future performance in terms of success and fail and grades of the students (Fernandes et al., 2019). Figure 35 presents the Math subject been used against all subjects for comparison and English subject been used against all subjects for comparison. English is one of the core subjects that is used to predict the performance of the students, with using the quizzes and monthly reports. English results were used for com parison with other subjects to estimate the future performance in terms of success and fail and grades of the students (Wang et al., 2018).

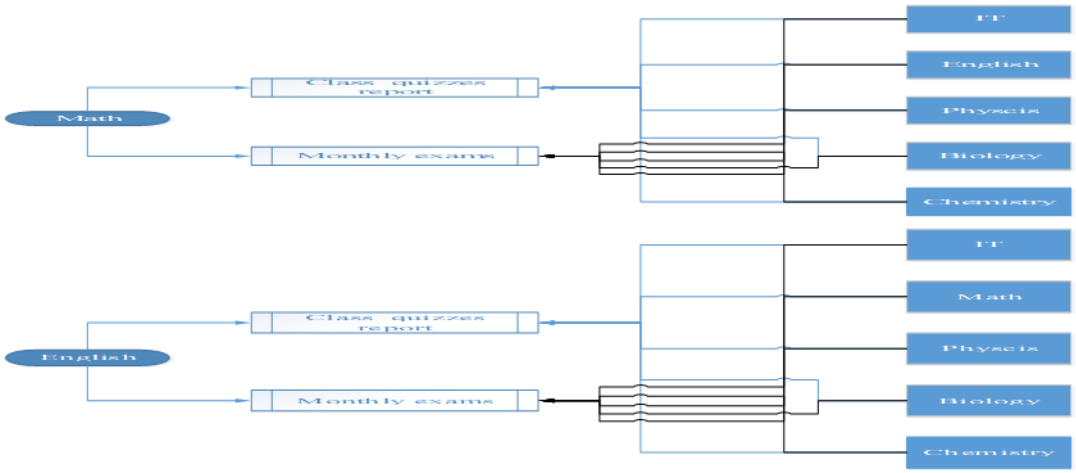

Figure 1. Math against all subjects and English against all subjects

Data Preparation: Student related data were collected from the four schools in Libya. High school student from science department were the sample of the study. 200 students were the data set for this research. The results of monthly exams and quizzes are the data to be used for the data mining to predict the performance (Menon & Islam, 2017). The subjects that are taken for the data mining model are divided into two categories. First category is two subjects that are Math and English.

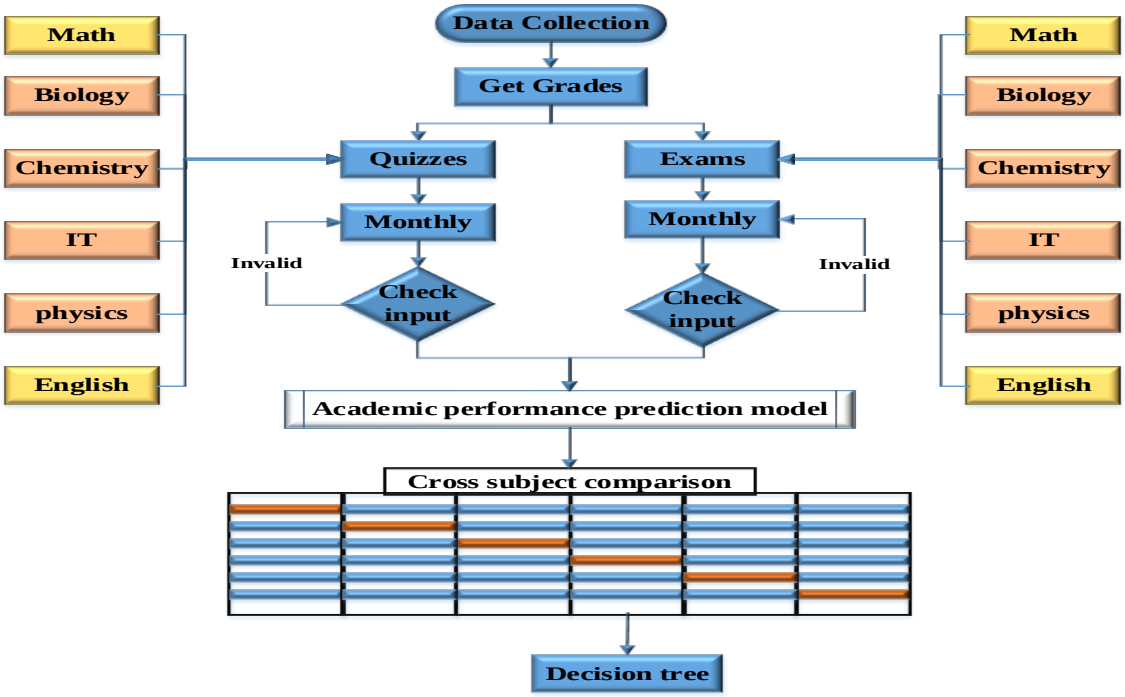

Figure 2. Presents the data processing(Rizvi, Rienties, Rogaten, & Kizilcec, 2020).

These are the core subjects that were used for the comparison and prediction. Second category is the rest of the main subjects that are (Biology, Chemistry, Physics, IT). These are the main and core subjects that decides the main performance of the students (Deeva & De Weerdt, 2018). Arabic and Islamic subjects are not taken into analysis because they do not perform onto the student's performance. The students were taken from different region to prevent bios and redundant results in

the analysis. Data Preprocessing: During data pre-processing, the student data from the main database where all the data of all subjects are stored. The data stored in several clusters, first cluster contains grades of quizzes of (physics, biology, IT, chemistry).

Second cluster contains grades of exams of (physics, biology, IT, chemistry). Third cluster contains grades of quizzes of (Math and English). Fourth cluster contains grades of exams of (Math and English) (Nebot, Mugica, & Castro, 2020). Teacher is responsible for the data collection and updates, the database contains records of all students, described by 6 parameters, including grades of quizzes of all (physics, biology, IT, chemistry), grades of quizzes of (Math and English), grades of exams of all (physics, biology, IT, chemistry), grades of exams of (Math and English), total score of all subjects (physics, biology, IT, chemistry), and total score of (Math and English). Essentially, the challenge in the presented data mining project is to predict the student future performance based on the collection of attributes providing information about the student previous results (Manjarres, Sandoval, & Suárez, 2018).

The selected target variable in this case, or the concept to be learned by data mining algorithm, is the "student final grade" (Romero & Ventura, 2006). A categorical target variable is constructed based on the original numeric parameter subject average score. It has five categories "excellent", "very good", "good", "average" and "fail". The five categories are determined from the total university score achieved by the students. A six-level scale is used in the Bulgarian educational system for evaluation of student performance at school (Castro, Vellido, Nebot, & Mugica, 2007). "Excellent" students are considered those who have a total score in the range between 5.50 and 6.00, "very good" in the range between 4.50 and 5.49, "good" in the range between 3.50 and 4.49, "average" in the range between 3.00 and 3.49, and "fail" in the range below 3.00.

DISCUSSION

Clustering analysis, for this research students are clustered based on their performance to find out the association rules that describe each cluster. The experiment was run three times on variation from 2 to 4 clusters. The evaluation of best number of clusters was based on the result of Euclidian distance between clusters. The best number of clusters was 3 clusters (Baroad, Hashim, Ahsan, & Zainal, 2020; Peral Cortés, Maté, & Marco Such, 2017). As the Simple K-Means clustering algorithm was used numeric values and can use nominal values by converting it to numeric values using optimal scale function. Subjects' marks percentage and final year results are chosen to apply clustering.

The analysis will be conducted in sections, first section is final exam grades of all subjects included in the test. Second section is quizzes grades of subjects. Next Association rules analysis that contains four sections (Aksoy, Narli, & Aksoy, 2020; M. M. Baroud, Hashim, Ahsan, Zainal, & Khalaff, 2020). First section is Math exam comparison against all subjects, second section is English exam comparison against all subjects, third section is Math quiz comparison against all subjects, fourth section is English quiz comparison against all subjects, following sections present the analysis.

Test Analysis Based on Final Grades

The following figures are a presentation of data for analysis based on subjects, the analysis was from high marks to low. The analysis was run from grade one to grade two(M. M. J. Baroud, Hashim, Zainal, & Ahnad, 2020; Yang & Li, 2018). The data is to be broken down according to the frequency occurrence. The grades were presented for the exams. All subjects are presented in the following figures.

English analysis: According to the results, majority of students have scored moderate grades. Low grades in this subject are very less compare to the Arabic subject. Even high score in this

subject is higher than Arabic. It looks like the teachers having hard time predicting and knowing how exactly the students will perform towards each subject. Lack of information and knowledge prevent the teachers from having the students perform well in the subjects.

Math analysis: Based results conducted from the analysis, majority of students still gathering around moderate grades. However low grades in Math subject are more than low grades in English and almost like Arabic subject. Moderate grades still are the most common grades among students.

Physics analysis: Based on the results from the analysis, unpredictably, most of the students in this subject have scored high scores, unlike other subjects. Should there be system where it records the student's performance, it should be easier to predict how the students will perform in each subject. The students interest data should be one the elements that the system takes into consideration during prediction. Physics subject is harder that Arabic and English, yet the students have performed well and scored higher. This means there must be some sort of the system that help the teachers to predicted and analysis the student's performance based on their background and interest and performance in the past years.

Chemistry analysis: based on the analysis conducted from the test, chemistry subject has gained more than majority of the student to score high grades. Low grades here are less than 20. Unlike English, Arabic, and Math. Chemistry is known to be a difficult subject and does not get most of the student's interest and understanding.

Biology analysis: Based on these results, the students' scores were majored between high and moderate, low grades are less than 10 and reaching to zero. This subject over all subjects is getting the best results. The students have better understanding of this subject more than other subjects. The reason for this success could be due to background, skills, interest, method of teaching. The teachers should find the student most interest and emphasis on it.

Information Technology analysis: The above results, present distinguished results as compared to the rest of the subjects, the students find this subject to be smoothed and easy to digest. None of the student have scored low mark. High grades were for most of the students. The massive of this subject is based on age, method of teaching, skills, and lifestyle of students. Teachers must take all these elements during analysis of student's performance. Fig 6 presents the results of information technology exam.

Summary of all subjects' exams in terms of grades: According to the above results, all the students can pass the subjects with moderate grades. The chances for low and fail grades are somewhat weak. Even though some subjects are difficult, yet the chance of passing at least with low is possible. The teachers must implement a systematic plan that ensures the performance of the students are up to standards. The systematic should be able to determine the weak areas of the students and work on improving them for better performance.

Total high scores of all subjects in terms of grades: These results are consistent with the previous results. As it presents that Information Technology is the most subjects that scored the highest grades among all subjects. It had attracted the attention of most of the students. Fig 8 presents over all high grades.

Total low scores of all subjects in terms of grades: The subject that received most low grades was Arabic subject as compared to all subjects. Even though this subject is the mother tongue for all students, yet majority of the students did not perform well. The reason of poor performance, the teachers did not teach well, students did not care about the material of the subject, teachers did not provide good materials, and poor teaching methods.

Total moderate scores of all subjects in terms of grades: According to these results, English and Biology have scored the most moderate grades among all subjects. These subject light subjects by nature that's why students are able digest and score moderate grades.

Quizzes Analysis Based on Final Grades

The following analysis is a presentation of data for analysis based on subjects, the analysis was from high marks to low. The analysis was run from grade one to grade two. The data is to be broken down according to the frequency occurrence. The grades were presented for the quizzes. following is the summery analysis of each subject. English analysis: Based on the results, majority of students still gathering around moderate grades. However low grades in Math subject are more than low grades in English and almost like Arabic subject. Moderate grades still are the most common grades among students.

Math analysis: Based on the results, majority of students still gathering around moderate grades. However low grades in Math subject are more than low grades in English and almost like Arabic subject. Moderate grades still are the most common grades among students.

Physics analysis: Based on the results, unpredictably, most of the students in this subject have scored high scores. Should there be system where it records the student's performance, it should be easier to predict how the students will perform in each subject. The students interest data should be one the elements that the system takes into consideration during prediction. Physics subject is harder than English, yet the students have performed well and scored higher. This means there must be some sort of the system that help the teachers to predicted and analysis the student's performance based on their past performance.

Chemistry analysis: According the results, chemistry subject has gained more than majority of the student to score high grades. Low grades here are less than 20. Unlike English and Math. Chemistry is known to be a difficult subject and does not get most of the student's interest and understanding.

Biology analysis: according to the results, biology has scored high marks. Majority of the students have gathered around high grades, as they have scored mostly moderate in math and high in English, this means students who perform well in math or English have the ability to perform well in biology based on the presented marks.

Information Technology analysis: according to the results, Information technology subject has scored moderate. Majority of the students have gathered around high and moderate grades, as they have scored mostly moderate in math and high in English, this means students who perform well in math or English can perform well in IT as well based on the presented marks.

Association Rules Analysis

The Apriori algorithm was used to find the rules that associate students' performance results. Science the algorithm takes only nominal or ordinal attributes; this research used (High, Moderate and Low) scale for all results as illustrated in section 4.2. To produce more accurate rules, the Apriori algorithm is applied on the clusters. The setting for algorithm was 0.1 for minimum support and 90% for confidence level is chosen this analysis focuses on the main factors that determine the performance of the students (Dubey, 2016). The main subjects that are used for this analysis are Math and English. They were compared with physics, biology, chemistry, and IT.

The reason of using these subjects and excluding Arabic and Islamic studies is these two subjects are not a factor to determine the students' performance and they considered secondary not primary subjects. Below is the analysis of the Math and English against other subjects. This analysis focuses on the main factors that determine the performance of the students. The main subjects that are used for this analysis are Math and English. They were compared with physics, biology, chemistry, and IT (Uswatun Khasanah, 2017). The reason of using these subjects and excluding Arabic and Islamic studies is these two subjects are not a factor to determine the students' performance and they considered secondary not primary subjects. Below is the analysis of the Math and English against other subjects.

Comparison Exam of Math Against other Subjects

Math VS Physics: The comparison was comparing the high, moderate, and low for both subjects to determine the student's performance. There is a possibility if students scored low marks in Math, they would score as well low marks in physics. Since Math is one of the core subjects and involved in physics calculation, it plays an important role in terms of prediction the student's performance. In this case it was obvious that those students who scored low marks, they did score almost same level in physics.

Math VS Chemistry: moderate and high grades are quite close in range. Math subject is somehow involved in chemistry as well. The level of difficulty is almost the same. If the student score low in the Math subject, there is a high chance that the student will score low grade as well in chemistry. As it was from the analysis, the moderate scores in Math are close to the high scores in chemistry. Which means that Math subject is a core subject to determine the performance of the students.

Math VS Biology: Math and biology are close in terms of moderate. Most of the grades are around moderate in both subjects. Though Math is not really part of biology, however students who are doing well in Math they can do well in biology. The difficulty of Math is higher than biology, so if students can handle difficult subject, it means they handle easy subject.

Math VS IT: most of the students scored moderate grades in Math also most of them scored high grades in IT. The IT subject could use Math in certain areas, therefore those students who understand Math subject, they do understand IT.

Math VS English: the English terms could be used in Math subjects. English is a general subject, if the student scored high grades along the year in Math, it means there is a high possibility that the student will score high grates in English as well.

Comparison Exam is English Against All Subjects

English VS Physics: English is the second core subject that is used to evaluate the student's current performance and predict their future performance in the exams for all subjects. Students for these two subjects have scored high and moderate grades. Low grades were very minimal.

English VS Chemistry: chemistry contains numerous English phrases. Most of the student who scored high in the chemistry subject have scored moderate in English and the scores were close. The number of students who scored high in chemistry and high in English is also close. This means that students who do chemistry, they can do well in English. Therefore, English is also used to determine the performance of the students because it can be used as a factor.

English VS Biology: students have scored high marks in both subjects. For majority of students, English is not an easy subject, in fact it is more difficult than biology. If students score high marks in English, it means it is easy for them to score high marks in biology or any theoretical subject.

English VS IT: students have scored high grades in IT and moderate grades in English. Table 10 presents English vs IT grades. IT and English share a lot of same phrases, therefore it is easy for students to manage these two subjects. If the students scored high grades in IT, it is easy for them to score either same grade or higher in English.

Comparison Quizzes is Math Against All Subjects

Math VS Physics: high grades of physics and moderate grades of Math which means that those students who scores low in physics, there is a chance they will score low in Math as well and vice versa. Since Math is one of the core subjects and involved in physics calculation, it plays an important role in terms of prediction the student's performance.

Math VS Biology: Math and biology are close in terms of moderate. Most of the grades are around moderate in both subjects. Though Math is not really part of biology, however students who are doing well in Math they can do well in biology. The difficulty of Math is higher than biology, so if students can handle difficult subject, it means they handle easy subject.

Math VS IT: most of the students scored moderate grades in Math also most of them scored high grades in IT. The IT subject could use Math in certain areas, therefore those students who understand Math subject, they do understand IT.

Math VS English: the grades are close in both subjects Math and English. Table 14 presents Math vs English grades. The English terms could be used in Math subjects. English is a general subject, if the student scored high grades along the year in Math, it means there is a high possibility that the student will score high grates in English as well.

Comparison Quizzes English Quizzes Against All Subjects

English VS Physics: English is the second core subject that is used to evaluate the student's current performance and predict their future performance in the exams for all subjects. Students for these two subjects have scored high and moderate grades. Low grades were very minimal.

English VS Chemistry: chemistry contains numerous English phrases. Most of the student who scored high in the chemistry subject have scored moderate in English and the scores were close. The number of students who scored high in chemistry and high in English is also close. This means that students who do chemistry, they can do well in English. Therefore, English is also used to determine the performance of the students because it can be used as a factor.

English VS Biology: students have scored high marks in both subjects. Table 18 presents English vs biology grades. For majority of students, English is not an easy subject, in fact it is more difficult than biology. If students score high marks in English, it means it is easy for them to score high marks in biology or any theoretical subject.

English VS IT: students have scored high grades in IT and moderate grades in English. Table 19 presents English vs IT grades. IT and English share a lot of same phrases, therefore it is easy for students to manage these two subjects. If the students scored high grades in IT, it is easy for them to score either same grade or higher in English.

According to the results observed in the analysis. The performance of the students was predicted and validated by using two core subjects, that are Math and English. It was clear that when students perform well in Math and English, there is a high chance for them to perform well in physics, biology, IT, and chemistry. Data mining is offering a great promise in organizing the student data and offering a great deal of a system that help the teacher to maintain the student's performance all over the semester from level to another level. Quizzes were analyzed same way as the exams.

Math results were compared against other subjects (English, Physics, IT, chemistry, Biology). The analysis was determined the possibility to predict the performance of the students based on the results of the Math quizzes results (Juhaňák, Zounek, & Rohlíková, 2019). Then English quizzes results were used for comparison against other subjects (English, Physics, IT, chemistry, Biology) to determine the possibility to predict the performance of the students based on the results of the English quizzes results.

The results determined the power of English and Math together to predict the student's performance in all subjects (Angeli, Howard, Ma, Yang, & Kirschner, 2017). Based on the above analysis, Math and English have an impact on the rest of the subjects. It was clear that students who scored high in Math or English they have scored either high or moderate in other subjects. This proves that Math and English are the most core subjects that can be used for future prediction in terms of student's performance (Burgos et al., 2018).

The algorithm has presented aa promising results (Romero & Ventura, 2006). Data mining database been used to assist the predicting process and to prepare the data set for modelling, to produce a clean data set, to prepare data for modelling process and to prepare a pattern of data to reach the prediction accuracy (Thakar, 2015).

It could help the teacher to create a teaching method that can benefit the students to perform better and to increase and sustain the number of student and at the same time can prevent them from leaving the school unexpected (Li, Gou, & Fan, 2019). It could help the management to find a way to reduce the number of potential students who are low achiever as early as possible. The result shows that logistic regression algorithm has a higher prediction accuracy (Jalota & Agrawal, 2019).

CONCLUSION

In this paper, it was discussed the ability to predict the student performance by using data mining technique. logistic regression algorithm has been used to determine the future performance by using their current and previous performance in the monthly exams and quizzes. The subjects were categorised into two categories. The Math and English were the main variables that are used in the prediction. CGPA was used to find out the final performance of the student. Predicting the performance and identifying the potential students at the early stage is very important for the management the predicting model should help to prevent the range of fail and low marks among the students.

ACKNOWLEDGMENT

We would like to thank my committee for supporting this research. Department of Information Technology and Engineering. Malaysia University of Science and Technology Kuala Lumpur, Malaysia

REFERENCES

Ahmed, A. M., Rizaner, A., & Ulusoy, A. H. (2016). Using data mining to predict instructor performance. *Procedia computer science, 102*, 137-142.

Aksoy, E., Narli, S., & Aksoy, M. A. (2020). An Educational Data Mining Application by Using Multiple Intelligences. In *Examining Multiple Intelligences and Digital Technologies for Enhanced Learning Opportunities* (pp. 93-110): IGI Global.

Angeli, C., Howard, S. K., Ma, J., Yang, J., & Kirschner, P. A. (2017). Data mining in educational technology classroom research: Can it make a contribution? *Computers & Education, 113*, 226-242.

Baroad, M. M., Hashim, S. Z. M., Ahsan, J. U., & Zainal, A. (2020). Positive region: An enhancement of partitioning attribute based rough set for categorical data. *Periodicals of Engineering and Natural Sciences (PEN), 8*(4), 2424-2439.

Baroud, M. M., Hashim, S. Z. M., Ahsan, J. U., Zainal, A., & Khalaff, H. (2020). Fast attribute selection based on the rough set boundary region. *Periodicals of Engineering and Natural Sciences (PEN), 8*(4), 2575-2587.

Baroud, M. M. J., Hashim, S. Z. M., Zainal, A., & Ahnad, J. (2020). *An New Algorithm-based Rough Set for Selecting Clustering Attribute in Categorical Data.* Paper presented at the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS).

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering, 66*, 541-556.

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In *Evolution of teaching and learning paradigms in intelligent environment* (pp. 183-221): Springer.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247-256.

Deeva, G., & De Weerdt, J. (2018). *Understanding automated feedback in learning processes by mining local patterns.* Paper presented at the International Conference on Business Process Management.

Dubey, R. (2016). Performance Evaluation Of Military Training Exercises Using Data Mining. In.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94*, 335-343.

Jalota, C., & Agrawal, R. (2019). *Analysis of educational data mining using classification.* Paper presented at the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

Juhaňák, L., Zounek, J., & Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior, 92*, 496-506.

Li, Y., Gou, J., & Fan, Z. (2019). Educational data mining for students' performance based on fuzzy C-means clustering. *The Journal of Engineering, 2019*(11), 8245-8250.

Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. S. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review*(33), 235-266.

Menon, A., & Islam, N. (2017). An investigation of the relationship between online activity on Studi. se and academic grades of newly arrived immigrant students: An application of educational data mining. In.

Nebot, À., Mugica, F., & Castro, F. (2020). An e-Learning Toolbox Based on Rule-Based Fuzzy Approaches. *Applied Sciences, 10*(19), 6804.

Peral Cortés, J., Maté, A., & Marco Such, M. (2017). Application of Data Mining techniques to identify relevant Key Performance Indicators.

Rizvi, S., Rienties, B., Rogaten, J., & Kizilcec, R. F. (2020). Investigating variation in learning processes in a FutureLearn MOOC. *Journal of computing in higher education, 32*(1), 162-181.

Romero, C., & Ventura, S. (2006). *Data mining in e-learning* (Vol. 4): WIT press.

Thakar, P. (2015). Performance analysis and prediction in educational data mining: A research travelogue. *arXiv preprint arXiv:1509.05176*.

Uswatun Khasanah, A. (2017). *A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques.* Paper presented at the Materials Science and Engineering Conference Series.

Wang, R., Ji, W., Liu, M., Wang, X., Weng, J., Deng, S., . . . Yuan, C.-a. (2018). Review on mining data from multiple data sources. *Pattern recognition letters, 109*, 120-128.

Yang, F., & Li, F. W. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education, 123*, 97-108.