Analyzing the Factors that Influence Enhancing Student Performance in Oman using Data Mining

Said Mohammed Alrashdi [1]

Akram Zeki [2]

[1] Department of Computer Science, International Islamic University Malaysia, KICT, Malaysia. Email: saidrashdi@unizwa.edu.om
[2] Department of Information Systems, International Islamic University Malaysia, KICT, Malaysia. Email: akramzeki@iium.edu.my

https://doi.org/10.20428/jst.v27i1.1981

Abstract:

Education field is a sign of advancement over the countries that can adopt technology to serve it. It will help to improve and enhance future achievements and be in touch with the development of technology utilizing solutions that extract student data, including their school records and other vital information about their performance, which can facilitate this process. These data are then analyzed to identify factors that affect the academic performance of the students at the school by expanding data mining techniques to enhance student academic performance. These factors are examined to develop a predictive model. Machine learning (ML) is one artificial intelligence (AI) field that can use such a model that supports educational institutions and decision-makers. A predictive method is applied using the data mining (DM) technique to take proactive action in identifying and anticipating the student's path. The data was analyzed, and the findings showed that the decision tree algorithm recorded the fastest training time for every 1000 rows. Also, the fast-scoring time for 1000 rows was in the decision tree algorithm, which was around 195 milliseconds, and the longest scoring time occurred in the random forest algorithm, which was two seconds. The top percent of classification errors reached 51% for the logistic regression algorithm and around +-1.5% of standard deviation. It took 520 mile-second for scoring time with 690 Gains for 67 m/s training time in every 1000 rows of the datasets. The findings of this study can help parents and teachers better understand the factors that influence students' academic performance and support them in assisting students with improving their academic performance.

Keywords: ML: Machine Learning, AI: Artificial Intelligence, DM: Data Mining, PSP: Predictive Student Performance

https://doi.org/10.20428/jst.v27i1.1981

1. Introduction

Recently, most of the contemporary applications of predicting the academic performance of secondary school students exploiting data mining techniques have become very important for improving students' performance. Its generated automatically using various information extraction or information integration approaches to take better action and straight decisions based on the relevant data. There are many factors for students at school, including general student information, education student information, behavior and activity, health status, and family affairs responsible for enhancing student performance. Using data science is very important for studying the relevant features that enhance students' future achievement in the education field and job seekers. One of these techniques is data mining because it supports analyzing factors that influence the students' academic performance at the school. Then, these factors are examined to develop a model that assists in enhancing the academic performance of the students [1] [2] [3]–[5].

Data mining technique is one of the data sciences that has become an effective solution for exploiting knowledge, ideas, experiences, and prediction skills. It helps to perform and enhance the student outcomes based on factors including general student information, education student information, behavior and activity, health status, and family affairs responsible for enhancing student performance while using this technique.

The modern approach to technology has become more effective in the presence of many modern and future technologies. That come related to each other such as the Internet of things, artificial intelligence, blockchain, cloud computing and the mechanism of providing security features, and data mining which naturally operate automatically without human intervention to pave the way for the creation and manufacture of digital content Integrated.

2. The Objectives and Scope of the Research

This study investigates the factors influencing students' academic performance at public secondary schools in Oman. Also the other object that can support the last goal is to propose a data mining-based approach to investigate the factors that enhance students' academic performance at the public secondary schools in Oman.

The scope of this research work is outlined in the following points:

- This study investigates the factors influencing students' academic performance at public secondary schools in Oman.
- This study uses data mining techniques to examine the factors influencing students' academic performance at public secondary schools in Oman.
- This study tries to identify the various factors that influence students' academic performance.

Figure 1 illustrates the research target view. It clarifies the path that will be followed in this research by knowing the scope of the study, which is the government schools for the secondary level in the Sultanate of Oman. The research will also focus on enhancing the student's academic performance during his/her study period and predicting the best outcomes in the future. According to the availability of a large amount of educational data related to the factors needed for this study, a data mining technique will be used in this research methodology. Therefore, ML is the technique for analyzing the datasets for classification in the part of prediction exploiting data mining technique.

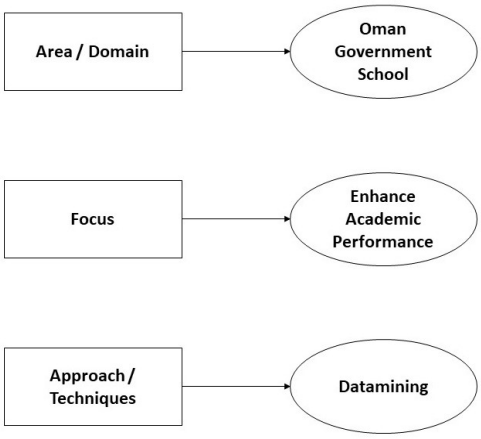

Figure 1: Research target view

https://doi.org/10.20428/jst.v27i1.1981

This study assists in improving the student's academic performance and helping their children achieve outstanding results at secondary schools in the Sultanate of Oman.

3. Predictive Student Performance Proposed Model

Students' demographic is a crucial factor for students' characteristics that include different type of data related to gender, age, salary income, and family background [6], [7]. External and environmental behaviors are also a part of factors coming from a student's life outside the class, including non-class activities, high school background for the parents, social interaction network, and health effects. It can raise the student's level and keep his/her study on a straight path [3], [4],[8], [9]. Activities that are not essential for normal classes and activities in preadmission courses for universities and schools give several reasons for predicting student's future goals in extracurricular activities, high school background factors[2]–[4], [10].

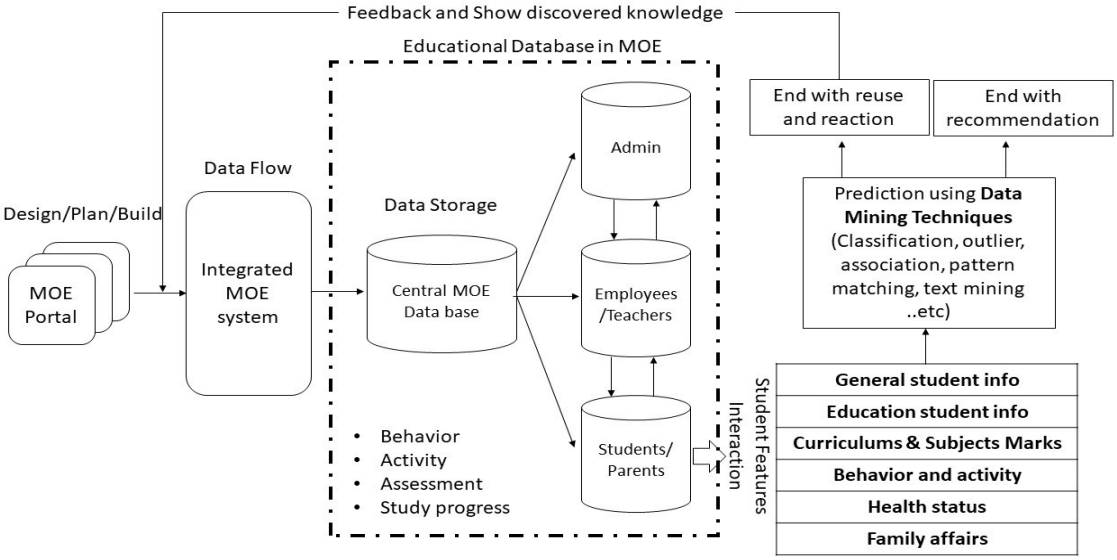

Figure 2: Proposed model (PSP) data mining in the portal of

Social interaction network is one of the educational factors that increase the feedback and reaction while interacting among students or teachers/instructors [11], [12]. One attribute can be collected after a student's enrolment at school or outside, such as attitude, motivation, personality, and learning strategies can deal with Psychometric factor[10], [13], [14].

The assessments for generating and finalizing study progress assist the students at schools in returning to his/her correct path by investigating the enhancement of students' performance. Figure 2 illustrates the proposed model for predictive student performance (PSP) using data mining for the MOE portal. Based on the factors to find a new source for an output, these sources or tools will assist in processing for improving the education system. Also, the process consumes time, cost, and effort to gather information to support students in making decisions and help them achieve their future ambitions and aspirations.

4. Methodology and Design

The proposed model consists of the three approaches (design, implementation, and evaluation) established in the design stages. It includes the approach for retrieving data from the MOE central database, estimating the values of the prediction performance, and assessing prediction quality.

The quantitative summary shows the performance of the proposed approach in terms of the total of the predictive comparisons used and the total execution time of the process through data mining technology to find the predictive results related to the factors affecting academic performance. Narrative synthesis is another way to report the findings from multiple experiments using selected words and text descriptions to summarize and explain the findings with comments. Furthermore, the proposed approach's performance for predicting the performance of the student's estimation using the data mining technique is evaluated by considering the most influential key factors in prediction values in the study environment, namely latency and accuracy.

https://doi.org/10.20428/jst.v27i1.1981

In addition, the accuracy of the proposed approach is evaluated with respect to the relative error between the current value and the future predictive performance values. That includes actual values and the prediction using the proposed approach that exploits the data mining technique.

A set of data was used to conduct various experiments related to the required predictions to know the evaluation path progress. The first set of experiments aims to investigate the impact of the size of the database quantity on the execution time that needs to predict the enhancing student values before and after data filtration.

Moreover, the second set of experiments attempts to examine the predictive rate's effectiveness on the value prediction quality during the data mining process. In contrast, the third set of experiments intends to evaluate the impact of the amount of dataset size on the relative error between actual values and predictive values. Furthermore, the fourth set of experiments aims to examine the impact of the user-given threshold value for the acceptable relative error between the real missing and the predictive values of enhancing student performance. Besides, the fifth set of experiments highlights studying the relationship between the expected factors of predicting the values of the student performance and the accuracy of the predicted values using the data mining technique. Lastly, the sixth set of experiments investigates the impact of the number of factors and batches on time latency to predict the values of enhancing student performance using the data mining technique. Finally, all experiment cases are considered with respect to the two existing types of datasets, predicted data value and actual datasets values.

The following figure 3 describes the detailed phases of the research methodology of this research work and the relevant activities flow with the progress to evaluate the proposed approaches.

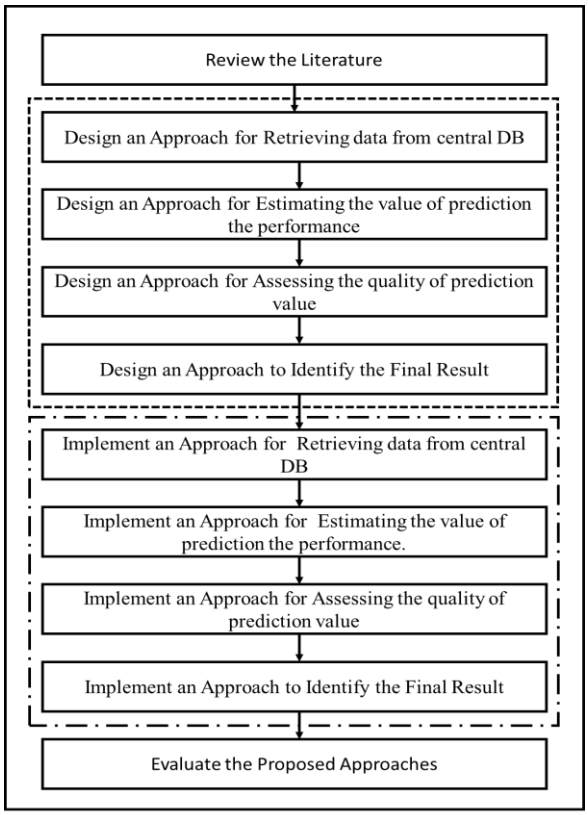

Figure 3: Research activities

5. Design and Implementation

The design and implementation will discuss the algorithm that will be used to automatically fetch the topics of the factors that influence student performance for secondary schools in the Sultanate of Oman. Also, it will cover an analysis of the factors that affect student performance to find the prediction in the mining as it is the scope of this research.

This analysis will define the algorithms based on the data mining technique using machine learning software that can provide real results based on the given data to support the decision-makers for acting. Then, this chapter will present the use of RapidMiner Studio as one of the machine learning software to retrieve and get the predictive algorithms. The RapidMiner Studio will be converted to a database table to retrieve the required information using SQL commands easily.

https://doi.org/10.20428/jst.v27i1.1981

The data relating to the research were obtained from the primary source in which Ministry of Education data are stored. It is real data for students studying in secondary schools in the Sultanate of Oman.

Around 40,000 records are available for this dataset that can provide a strong result and prepare a standard step for design and implementation using machine learning software to analyze the requirements. The dataset's attributes are divided into three parts (Student, Teacher, and Parents) because it brings different factors that can affect students' performance at secondary schools in the Sultanate of Oman using data mining techniques. After the filtration of data, around 8000 records and more than 22 attributes within student datasets.

Table 1: Table of Predictive Student Performance Group

cluster *Category*	Age_std *Number*	FamilyIncome *Number*	Gender *Category*	Willaya *Category*	التربية الاسلامية *Number*	الدراسات الاجتماعية *Number*	الرياضيات *Number*	اللغة الانجليزية *Number*	اللغة العربية *Number*
PSP_0	21.000	0	ذكر	سمائل	61.000	72.000	50.000	66.000	62.000
PSP_4	18.000	0	أنثى	أدم	61.000	56.000	80.000	80.000	63.000
PSP_1	16.000	0	ذكر	ازكي	61.000	74.000	78.000	89.000	78.000
PSP_0	17.000	80.000	ذكر	بهلاء	61.000	50.000	38.000	30.000	80.000
PSP_0	20.000	100.000	ذكر	منح	61.000	50.000	58.000	28.000	82.000
PSP_4	17.000	120.000	ذكر	بهلاء	61.000	65.000	75.000	84.000	68.000
PSP_0	19.000	135.000	ذكر	بدبد	61.000	50.000	30.000	72.000	50.000

Table 1 shows the process of finding a group of predictive student performances that can make it a feature selection for the datasets to find the factors influencing student performance in grade 10 at secondary schools in Oman.

6. Results of ML Outcomes

This part covers an analysis of the factors that affect student performance to find the prediction in the mining as it is the scope of this research. This analysis will define the algorithms based on the data mining technique using machine learning software that can provide actual results based on the given data to support the decision-makers for acting.

Then, this will present RapidMiner Studio as a machine learning software to retrieve and get the optimal predictive algorithm. The RapidMiner Studio will be converted to a database table to retrieve the required information using SQL commands easily.

ML is a trusted, referenced, and accountable technique when it comes to predicting student performance data analytics. Many applications have worked as machine learning in analytics when they want to deal with tons of data to get the predictive outcomes that help to get new insight into a decision and future needs to understand how to use ML. Figure 4 illustrates the relationships between AI and ML and how it can be essential to perform the factors that affect student performance and get the predictive outcomes.

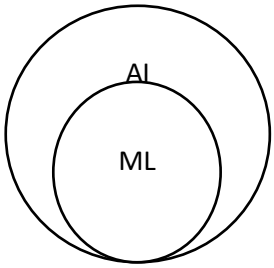

Figure 4: AI and ML relation

The algorithms and mathematics series are the methods in data mining for getting the predictive outputs based on the giving historical data for factors that affect the student performance to recognize faces better and make the changing of those predictive issues.

Machine learning models the data for predicting students' performance at schools in the Sultanate of Oman by comparing the algorithms and finding the most relevant for our study. Table 2 shows the outcomes of the data selection process and the proper algorithm tested for 168 models. The fast significant margin classification error is 24.40%, and the standard deviation is 0.5%, the lowest percentage among other algorithms measurements.

https://doi.org/10.20428/jst.v27i1.1981

Also, the data is analyzed, and the finding of the fastest training time for every 1000 rows occurs in the decision tree algorithm. Around 195 milliseconds is the fast-scoring time for 1000 rows in the decision tree algorithm, and the longest scoring time occurs in the process for the random forest algorithm for 2 seconds.

Figure 5 visualizes the percentage of classification errors for extracting the predictive models/algorithms. The top percent of classification errors reached 51% for Logistic Regression with around +-1.5% of Standard Deviation, and it took 520 mile-second for scoring time with 690 Gains for 67 m/s training time in every 1000 rows. The minimum time recorded for this training in the decision tree model registered 195 mile-second scoring time for every 1000 rows, and the classification error percentage was nearly 31% within +-2.1% of the standard deviation.

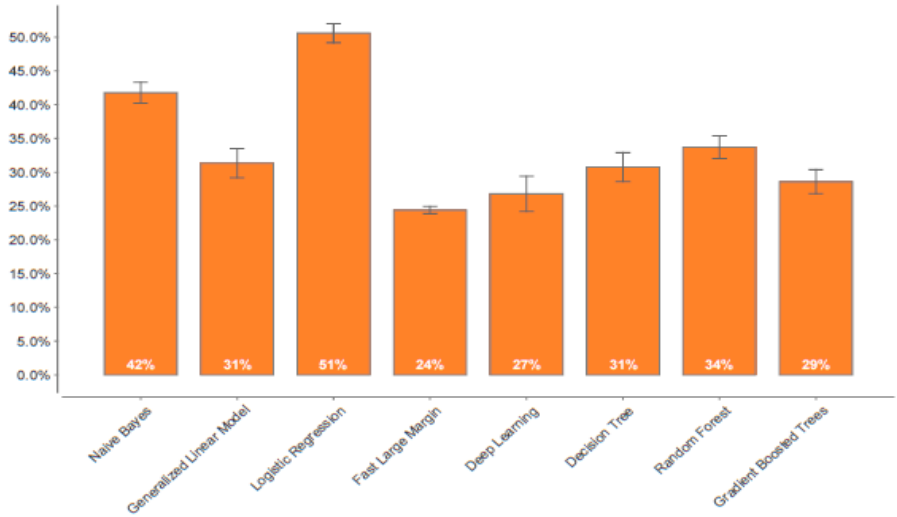

Figure 5: Classification error chart view

Table 2: Classification Error of Algorithms

Model	Classification Error	Standard Deviation	Gains	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Naïve Bayes	41.80%	+-1.6%	1058	1 min 6 s	34 ms	488 ms
Generalized Linear Model	31.30%	+-2.1%	1498	55 s	555 ms	353 ms
Logistic Regression	50.60%	+-1.4%	690	2 min 31 s	67 ms	520 ms
Fast Large Margin	24.40%	+-0.5%	1796	4 min 14 s	1 s	1 s
Deep Learning	26.80%	+-2.6%	1686	2 min 48 s	1 s	539 ms
Decision Tree	30.70%	+-2.1%	1522	1 min 3 s	16 ms	195 ms
Random Forest	33.70%	+-1.7%	1396	15 min 13 s	100 ms	2 s
Gradient Boosted Trees	28.60%	+-1.8%	1614	7 min 35 s	657 ms	989 ms

7. Research Area and Direction

The researcher has presented the design and the implementation of the algorithms that are intended to generate the topics of predictive student performance using data mining techniques. The approach attempts to determine the mining algorithms while using the updated database to find the prediction for future vision. Moreover, these algorithms were implemented to affect students' performance at secondary schools. The phases of the proposed approach are discussed with a running database example to explain the detailed steps of each phase. Therefore, the implementation was useful in finding the results of all performance datasets to act and make a decision.

8. Conclusion and Recommendation

The predictive method is a data mining technique based on the principle of machine learning in a way that simulates artificial intelligence. The education field is an important place for using real data to find future achievements and assist the decision-makers in taking action by processing the data to get the outcomes. This research focuses on the factors that enhance student performance at secondary schools in the Sultanate of Oman by exploiting data mining to get the classification method for predicting performance during the study period.

https://doi.org/10.20428/jst.v27i1.1981

Six predictive algorithms were found using machine learning while preparing the classification process. The top percent of classification errors reached 51% for the logistic regression algorithm and around +-1.5% of standard deviation. It took 520 mile-second for scoring time with 690 Gains for 67 m/s training time in every 1000 rows of the datasets. The data was analyzed, and the decision tree algorithm recorded the fastest training time for every 1000 rows. Also, around 195 millisecond is the fast-scoring time for 1000 rows in the decision tree algorithm and the longest scoring time occur in the process for the random forest algorithm for 2 seconds. The findings of this study can help parents and teachers to better understand the factors that influence students' academic performance and support them in assisting students with improving their academic performance.

The use of machine learning technology through the presence of real data is one of the techniques which can extrapolate a set of future predictions and provide the necessary requirements and needs for improvement. It contributes to determining the appropriate period according to the cumulative data that derives its updates continuously, consecutively, and steadily over long periods using data mining technology.

9. References

[1] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. & Educ., vol. 113, pp. 177–194, 2017, doi: 10.1016/j.compedu.2017.05.007.

[2] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," 2014 International Conference on Communication and Network Technologies. IEEE, 2014. doi: 10.1109/cnt.2014.7062736.

[3] D. M. D. Angeline, "Association Rule Generation for Student Performance Analysis using Apriori Algorithm," SIJ Trans. Comput. Sci. Eng. & its Appl., vol. 01, no. 01, pp. 16–20, 2013, doi: 10.9756/sijcsea/v1i1/01010252.

[4] S. Natek and M. Zwilling, "Student data mining solution–knowledge management system related to higher education institutions," Expert Syst. Appl., vol. 41, no. 14, pp. 6400–6407, 2014, doi: 10.1016/j.eswa.2014.04.024.

[5] Z. K. Papamitsiou, V. Terzis, and A. A. Economides, "Temporal learning analytics for computer based testing," Proceedings of the Fourth International Conference on Learning Analytics And Knowledge. ACM, 2014. doi: 10.1145/2567574.2567609.

[6] T. M. Christian and M. Ayub, "Exploration of classification using NBTree for predicting students' performance," 2014 International Conference on Data and Software Engineering (ICODSE). IEEE, 2014. doi: 10.1109/icodse.2014.7062654.

[7] U. bin Mat, N. Buniyamin, P. M. Arsad, and R. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," 2013 IEEE 5th Conference on Engineering Education (ICEED). IEEE, 2013. doi: 10.1109/iceed.2013.6908316.

[8] P. M. Arsad, N. Buniyamin, and J. A. Manan, "A neural network students' performance prediction model (NNSPPM)," 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA). IEEE, 2013. doi: 10.1109/icsima.2013.6717966.

[9] S. Parack, Z. Zahid, and F. Merchant, "Application of data mining in educational databases for predicting academic trends and patterns," 2012 IEEE International Conference on Technology Enhanced Education (ICTEE). IEEE, 2012. doi: 10.1109/ictee.2012.6208617.

[10] T. Mishra, D. Kumar, and S. Gupta, "Mining Students' Data for Prediction Performance," 2014 Fourth International Conference on Advanced Computing & Communication Technologies. IEEE, 2014. doi: 10.1109/acct.2014.105.

[11] C. Tucker, B. Pursel, and A. Divinsky, "Mining Student-Generated Textual Data In MOOCS and Quantifying Their Effects on Student Performance and Learning Outcomes," 2014 ASEE Annual Conference & Exposition Proceedings. ASEE Conferences. doi: 10.18260/1-2--22840.

[12] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," Comput. & Educ., vol. 68, pp. 458–472, 2013, doi: 10.1016/j.compedu.2013.06.009.

[13] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," 2014 IEEE International Advance Computing Conference (IACC). IEEE, 2014. doi: 10.1109/iadcc.2014.6779384.

[14] I. Hidayah, A. E. Permanasari, and N. Ratwastuti, "Student classification for academic performance prediction using neuro fuzzy in a conventional classroom," 2013 International Conference on Information Technology and Electrical Engineering (ICITEE). IEEE, 2013. doi: 10.1109/iciteed.2013.6676242.