ICIC Express Letters
Part B: Applications
Volume **10**, Number **8**, August **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **681–687**

A STUDY ON ADOPTION OF DATA MINING TECHNIQUES TO ANALYZE ACADEMIC PERFORMANCE

Hussah Talal[1] and Saqib Saeed[2]

[1]Department of Computer Science
[2]Department of Computer Information Systems
Imam Abdulrahman Bin Faisal University
P.O. Box 1982, Dammam 31441, Saudi Arabia
{ 2190500050; sbsaed }@iau.edu.sa

Received January 2019; accepted April 2019

Abstract. *Extensive availability of academic data requires appropriate structuring and analysis applications to generate knowledge which can be used to improve the academic process. Data mining classification and clustering algorithms have huge potential to benefit this sector. In this paper, we have applied different classification and clustering algorithms on an educational dataset by using WEKA. The results highlighted that these algorithms can effectively predict the student performance after training; however, selection of an appropriate algorithm is very critical. We urge the academicians and educational administrators to use different data mining applications to track and predict student academic performance and to use this knowledge as a baseline to design policies and guidelines for the better learning experience of students in academic settings.*
Keywords: Educational dataset, Education, k-nearest neighbor, J48 decision trees, ZeroR algorithm, Clustering, k-means

1. **Introduction.** Education is a crucial sector not only for human development but also for the economic uplift of any society. Therefore, there is a lot of emphasis by every government to improve this sector and recent technological innovations have shown huge promise for the improvement of the education sector. Data mining is one such technology that can be applied on academic data to chalk out optimal strategies [1]. Arnold developed an algorithm for providing an early warning system for the weak student to improve their academic performance [2]. Cortez and Silva applied three different data mining techniques on secondary school dataset to study their performance [3]. Okfalisa et al. carried out a comparative study between k-nearest neighbor and modeled k-nearest neighbor algorithm and they found modeled k-nearest neighbor algorithm more accurate [4]. Similarly, Ozer carried out a comparative study on large dataset to analyze the performance of different algorithms [5]. Romero and Ventura have emphasized the need for application of data mining techniques in various knowledge areas to improve their effectiveness and highlighted huge potential of data mining in the education sector [6]. Fernandes et al. carried out a study to predict students' performance based on their analysis of students record in public schools of Brazil using classification models based on the Gradient Boosting Machine. They found that not only academic performance but also students demographic characteristics help in improving the accuracy of prediction [7]. Ahuja et al. highlighted that educational data mining can help in improving student's effectiveness by analyzing information stored in large database systems managed by the educational institutions [8]. Kumar et al., used visualizations to better understand the relationships in the educational dataset [9]. Alom and Courtney highlighted that there is a need for more detailed studies to optimally use educational data mining in forecasting

DOI: 10.24507/icicelb.10.08.681

students' academic performance [10]. Similarly, Shingari and Kumar argued that there is a huge promise in implementing data mining applications in the educational setting, and more rigorous studies are required to enrich this body of knowledge [11].

Keeping this in view, in this paper we are presenting another case study to further enrich this body of knowledge. The results of this study will help in realizing more the benefits of adopting data mining techniques in the education sector. The process of data mining is shown in Figure 1. Our intention is that such analysis and future predictions of students' performance are helpful for educational institutions to monitor the progress of their students. The predictions can help educational institutions to better design their organizational processes to improve the academic performance of students.

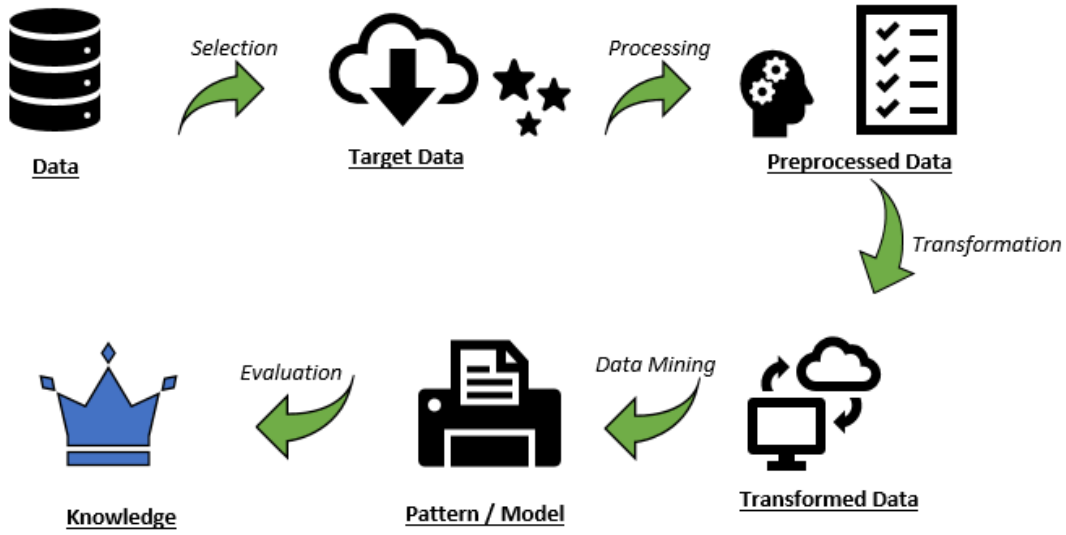

FIGURE 1. Data mining process

The rest of the paper is structured as follows. Section 2 discusses the methodology employed to generate results followed by findings in Section 3. Section 4 provides a discussion followed by a conclusion in Section 5.

2. **Methodology.** The nature of our study is exploratory, highlighting how data mining techniques can be used to analyze and understand students' performance and to predict future results based on historical data. We used secondary data for our analysis which was Students' Academic Performance Dataset [12]. The reason for selecting this dataset was based on data currency and conveyance. The dataset attributes are shown in Table 1.

Since we were mainly looking at the academic performance, during preprocessing stage, we removed Nationality, Place of birth, Educational stages, Grade levels, Topic, Semester, Parent responsible for student attributes to optimize the response time. We used experience API (xAPI) tool for learner activity tracker. Before conducting the experiment, we removed the data from "class" column, and we used our result to compare this column with our experimental results. In order to conduct the experiment, we used the Waikato Environment for Knowledge Analysis (WEKA) software of machine learning to apply techniques of data mining on our dataset and trained the models [13]. In order to fully understand the effectiveness of data mining capabilities, we applied both classification and clustering data mining techniques. The reason for applying both techniques was to understand fully the effectiveness of both approaches in different scenarios. For classification data mining approach, we used the k-nearest neighbor algorithm, j48 decision trees

Table 1. Dataset attributes

Attribute	Description
Gender	Student gender: male – female
Nationality	Nationality of students: Kuwait – Lebanon – Egypt – Saudi Arabia – USA – Jordan – Venezuela – Iran – Tunis – Morocco – Syria – Palestine – Iraq – Lybia
Place of birth	Place of birth of students: Kuwait – Lebanon – Egypt – Saudi Arabia – USA – Jordan – Venezuela – Iran – Tunis – Morocco – Syria – Palestine – Iraq – Lybia
Educational stages	Student educational level: Lower level – Middle School – High School
Grade levels	Student grade G-01 – G-02 – G-03 – G-04 – G-05 – G-06 – G-07 – G-08 – G-09 – G-10 – G-11 – G-12
Section ID	Student classroom: A, B, C
Topic	Topic of Courses: English – Spanish – French, Arabic – IT – Math – Chemistry – Biology – Science – History – Quran – Geology
Semester	School year semester: First – Second
Parent responsible for student	Mom – father
Raised hand	Number of times a student raised his/her hand in the classroom
Visited resources	Number of times a student visited the resources
Viewing announcements	Number of times a student viewed the new announcements
Discussion groups	Number of times a student participated in discussion groups
Parent answering survey	Has parent answered school surveys or not: YES – NO
Parent school satisfaction	Satisfaction of the parent from school: YES – NO
Student absence days	Above-7 – Under-7
Class	Determine the level of the student depending on her/his grades: Low-Level (L): degree (from 0 to 69) – Middle-Level (M): degree (from 70 to 89) – High-Level (H): degree (from 90 to 100)

algorithm, random forest algorithm, and ZeroR algorithm. The reason for selecting these algorithms was their effectiveness highlighted in data mining literature. In order to test the effectiveness of each algorithm, we compared the predicted results of each algorithm about attributes such as raised hand, visited resources, viewing announcements, discussion groups, parent answering survey, parent school satisfaction, and student absence days. Furthermore, we used k-means algorithm on the data set, which is highlighted as the most effective clustering technique in the data mining literature.

3. **Findings and Results.** In this section, we discuss the findings of our experiment.

3.1. **Classification algorithms.** In the first phase of our experiment, we applied k-nearest neighbor, J48 decision trees, random forest, and ZeroR algorithms on our dataset. The predicted results were compared with the original results of the dataset. We used the following attributes in our dataset: gender, raised hand, visited resources, viewing announcements, discussion groups, parent answering survey, parent school satisfaction, student absence days, and class. Our focus was on the CLASS column (this column displays the level of the student based on his/her previous behavior in the educational environment), where we trained our models using different algorithms to predict the

level of the student performance. Then, we compared the accuracy of these algorithms in predicted data. We also have compared between them using other characteristics that we explained later in this research.

Firstly, in WEKA, we applied k-nearest neighbor algorithm to our dataset and it took 0.12 seconds to test the model on the training data. All the 480 instances of dataset were correctly classified by the algorithm. As shown in Figure 2, this algorithm achieved 100 percent accuracy and all the three class levels were correctly identified as per the original dataset. This meant that k-nearest neighbor algorithm is very efficient, accurate and suitable for applying on students' performance dataset.

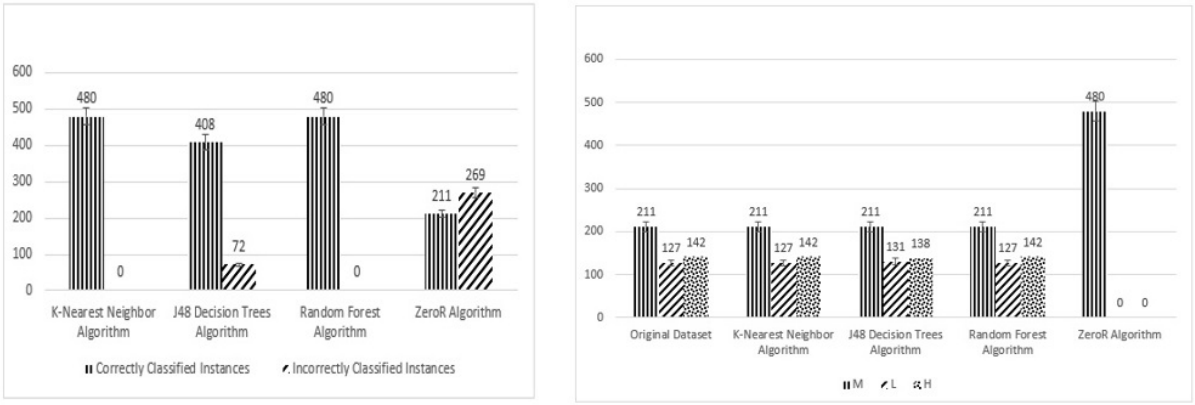

FIGURE 2. Performance results of four algorithms

Secondly, we applied j48 decision trees algorithm on the dataset. It took 0.08 seconds to build the model and 0.04 seconds to test the model. As shown in Figure 2, 85% instances were correctly intensified.

Thirdly, we applied the random forest algorithm on the students' performance dataset and it took 0.48 seconds to build the model and 0.09 seconds to test the model on training data. As Figure 2, it highlights this algorithm also achieved 100 percent accuracy, as prediction results matched exactly with the original dataset. This highlights that the random forest algorithm is also very effective while applying on students' performance dataset.

Lastly, we applied ZeroR algorithm on the students' performance dataset and the algorithm returned all the records as the middle level which is not accurate and out of 480 instances, only 211 were termed as correct. This highlighted that using this algorithm on the dataset resulted in incorrect predictions.

3.2. **Clustering data algorithms.** In the second phase of the experiment we applied k-means clustering algorithm on our dataset. We used the following attributes in our dataset: raised hand, visited resources, viewing announcements, discussion groups, parent answering survey, parent school satisfaction, student absence days, and class. It took 0.04 seconds to build the model of the data. Table 2 highlights the relationship between three clusters and level of students.

TABLE 2. Clustered data

Cluster	Data in cluster	Number	Percentage
0	90, 90, 57, 22, Yes, Good, Under-7, H	173	36%
1	92, 65, 62, 53, Yes, Good, Under-7, H	142	30%
2	20, 12, 15, 70, No, Good, Above-7, L	165	34%

The graphical distribution of data is shown in Figure 3. We applied the k-means algorithm and fixed the value of k at 3. The algorithm analyzed the performance data and clustered based on the relationships to make it more understandable.

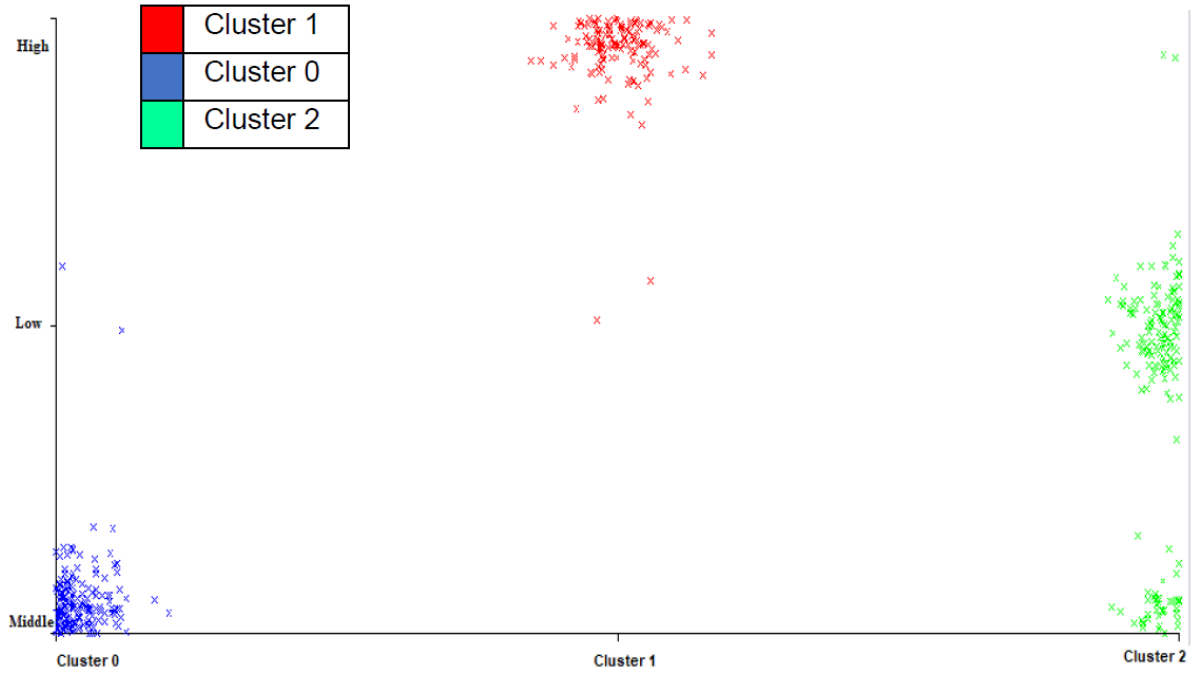

FIGURE 3. Distribution of data in clusters

Table 3 highlights the results from the WEKA, and this helps in determining the common relationships in data. As attributes highlight that there is a relationship between parent satisfaction, survey response and students' academic performance, we can easily visualize different sets of data based on the relationships among data. Cluster 2 students have a bad performance, cluster 0 students have average response whereas cluster 1 students have good academic performance.

TABLE 3. Cluster relationships

Attribute	Full data	Cluster 0	Cluster 1	Cluster 2
Raised hands	46.775	52.9653	70.7324	19.6667
Visited resources	54.7979	67.9653	79.1549	20.0303
Announcements view	37.9188	43.6821	53.9437	18.0848
Discussion	43.2833	43.8324	53.6127	33.8182
Parent answering survey	Yes	Yes	Yes	No
Parent school satisfaction	Good	Good	Good	Bad
Student absence days	Under-7	Under-7	Under-7	Above-7
Class	M	M	H	L

4. **Discussion.** In order to acquire optimal prediction results, selection of an appropriate algorithm is the most critical step. The correct choice of the algorithm leads to the accurate model to predict the necessary actions required to raise the academic standard of students and design of proactive strategies to solve various problems that arise in the academic environment. The accuracy and the effectiveness of the algorithms are affected also by the dataset characteristics so this study highlights effectiveness of different algorithms. In order to understand the effectiveness of data mining in educational data, we applied

four prediction techniques on our dataset by first training them and later comparing predictions with actual results. The results of the prediction process are different depending on the classification data mining algorithms being employed on the dataset. Based on this we can deduce that k-nearest neighbor and random forest algorithms achieved 100 percent accuracy. Therefore, using these algorithms on our dataset is very useful. The results showed that j48 decision trees algorithm yielded 85% of accuracy which is less than the k-nearest neighbor algorithm and the random forest algorithm. As shown in Figure 4, the performance of ZeroR algorithm was poor compared with all others which achieved 44% accuracy on our dataset, so using ZeroR algorithm is not recommended for similar datasets.

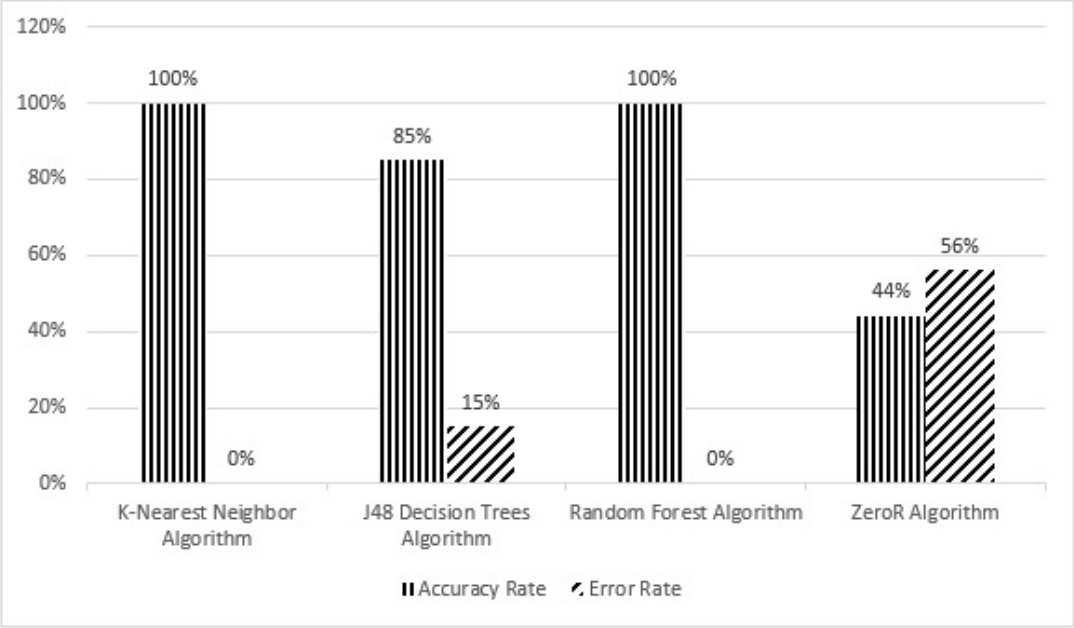

FIGURE 4. Comparative performance of classification algorithms

Furthermore, in situations where a very large amount of random data is available, and we are interested to classify the data to get better insights of the data, clustering is proved to be an effective mechanism. Clustering data mining algorithms group the related data in one cluster and find the arithmetic average. Then, the elements in the group are closer to the arithmetic mean of this group than the arithmetic average of any other group [12]. This helps to deal with a large amount of data by summarizing a vast amount of data into few categories in order to facilitate its analysis. In this research, we used the k-means clustering data mining technique, which divided the data into three clusters which made it easier to understand the data and the relations between them. This clustering helped to identify the main reasons for the problems that were related to the performance of students in the dataset. This information can help in improving educational administration in the educational institutes. Especially as many educational institutions use virtual learning environments such as blackboards to interact with students resulting in the availability of large data. Data mining applications can be applied to this data to improve students learning and understand their behavior. The predictions can be used to design appropriate policies and remedial actions to solve recurring problems. This will help in fostering conducive student-centric learning environments in educational institutions.

5. **Conclusion.** Data mining applications can be effectively used in the education sector as well, to analyze the student performance and to better design counseling mechanisms to improve student performance. The relationship analysis can help to identify the patterns

in underlying data and barriers in fostering better learning experiences for students. This is exploratory research where we applied four classification and one clustering algorithm on an educational dataset. WEKA analysis highlighted that k-nearest neighbor algorithm and random forest algorithm showed a hundred percent accuracy. k-means clustering algorithm helped to classify dataset into different clusters to better understand the underlying constructs and the relationships among different clusters. This research emphasized the effectiveness of data mining algorithms on the educational dataset, so the findings provide a positive sign for academicians and administrators to use such technologies in their organizational settings to improve the student performance. There is a need by the scientific community to design easy to use applications which can be used by academic administrators in improving the learning environments in their institutions. As future work, these algorithms can be deployed in practice on live data in different academic settings to improve learning in practice.

REFERENCES

[1] P. Kamal and S. Ahuja, Academic performance prediction using data mining techniques: Identification of influential factors effecting the academic performance in undergrad professional course, in *Harmony Search and Nature Inspired Optimization Algorithms*, N. Yadav, A. Yadav, J. C. Bansal, K. Deep and J. H. Kim (eds.), Singapore, Springer, 2019.

[2] K. E. Arnold, Signals: Applying academic analytics, *Educause Quarterly*, vol.33, no.1, 2010.

[3] P. Cortez and A. Silva, *Using Data Mining to Predict Secondary School Student Performance*, EUROSIS, 2008.

[4] Okfalisa, I. Gazalba, M. Mustakim and N. G. I. Reza, Comparative analysis of k-nearest neighbor and modeled k-nearest neighbor algorithm for data classification, *The 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering*, 2017.

[5] P. Ozer, *Data Mining Algorithms for Classification*, BSc Thesis, Radbound University Nijmegen, http://www.socsci.ru.nl/idak/teaching/batheses/bachelor_thesis_patrick_ozer.pdf, 2008.

[6] C. Romero and S. Ventura, Data mining in education, *Wiley Int. Rev. Data Min. and Knowl. Disc.*, vol.3, no.1, 2013.

[7] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho and G. van Erven, Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil, *Journal of Business Research*, vol.94, pp.335-343, 2019.

[8] R. Ahuja, A. Jha, R. Maurya and R. Srivastava, Analysis of educational data mining, in *Harmony Search and Nature Inspired Optimization Algorithms*, N. Yadav, A. Yadav, J. C. Bansal, K. Deep and J. H. Kim (eds.), Singapore, Springer, 2019.

[9] J. D. Kumar, K. R. Shankar and R. A. K. Saravanaguru, An investigation on educational data mining to analyze and predict the student's academic performance using visualization, in *Information Systems Design and Intelligent Applications*, S. C. Satapathy, V. Bhateja, R. Somanah, X.-S. Yang and R. Senkerik (eds.), Singapore, Springer, 2019.

[10] B. M. M. Alom and M. Courtney, Educational data mining: A case study perspectives from primary to university education in Australia, *International Journal of Information Technology and Computer Science*, vol.2, no.2, pp.1-9, 2018.

[11] I. Shingari and D. Kumar, *A Survey on Various Aspects of Education Data Mining in Predicting Student Performance*, JASC, 2018.

[12] E. A. Amrieh, T. Hamtini and I. Aljarah, Mining educational data to predict student's academic performance using ensemble methods, *International Journal of Database Theory and Application*, vol.9, no.8, pp.119-136, 2016.

[13] *Weka 3: Data Mining Software in Java*, https://www.cs.waikato.ac.nz/ml/weka/, 2018.