Using Data Mining Techniques in Building a Model to Determine the Factors Affecting Academic Data for Undergraduate Students

Faisal Mohammed Nafie[1] and Abdelmoneim Ali Mohamed Hamed[2]

fm.ali@mu.edu.sa aa.mohamed@mu.edu.sa

- [1]Department of Computer Science, College of Science and Humanities at Alghat, Majmaah University, Al-Majmaah, 11952, Saudi Arabia.
- [2]Department of Mathematics, College of Science and Humanities at Alghat, Majmaah University, Al-Majmaah 11952, Saudi Arabia.

Abstract

The main goal of higher education institutions is to present a high level of quality education to its students. This study uses data mining techniques to extract educational data from cumulative databases and used them to make the right decisions. This paper also aims to find the factors affecting students' academic performance in Majmaah University, KSA, during 2010 - 2017 period. The study utilized a sample of 6,158 students enrolled from two colleges, males and females. The results showed a high percentage of stumbling and dismissed between graduate and regular students where more than 62.5% failed to follow the plan. Only 2% of students scored distinction during their study of all graduated since their grade point average, secondary level, was statistically significant, where $p<0.05$. Dismissed percentage was higher among males. These results promoted some recommendations in which decision-makers could take them in considerations for better improvement of academic achievements: including of specialized programs to follow-up in regards to stumbling and failure. Utilization of different communication tools are needed to activate academic advisory for dismiss and dropout evaluation.

Keywords: *Academic performance; Data Mining; Higher education; Academic advising; Decision-making.*

1. Introduction

The availability of data stored in the databases in large quantities encouraged researchers to take advantage of them in many vital areas. In Recently years many Universities and Institutes in higher education have been concerned about the quality education and use different ways to analyse and improve the understanding of students success and academic achievement. Data mining is the process of exploring data from different perspectives and summarizing it into beneficial information (Ali & Tuteja 2014). Data mining techniques can be used to extract information and discover knowledge from it to solve problems and make decisions (Abu-Oda & El-Halees 2015;Guleria & Sood 2014;Aslam & Ashraf 2014). It refers to Knowledge discovery in databases (KDD) that covers a complex process from data preparation to knowledge modelling (Jain, et. al, 2017). It predicts the failure of students in real-time data from school or graduate students and to detect the failure of students to improve their academic performance and prevent them from dropping out (Khobragade & Mahadik 2015). The main function of Data Mining techniques is using various methods and algorithms to find and mine patterns of stored data (Suhirman, et. al, 2014; Rupali 2013). Therefore, prospecting techniques are used to support higher education institutions to make better decisions, to plan and develop advanced follow-up for students, to predict their academic achievement with greater accuracy (Shruth & Chaitra 2016), and to enable institutions and organizations to use resources and human personnel effectively (Agnihotri & Mishra 2015). Universities and Institutes may trail the best of the educations; but still, they suffer the problem of dropout students, low achievers and jobless students (Thakar 2015).

This work was applied for Majmaah University in KSA. It has a huge database containing academic data for students, academic programs, courses, results, educational outputs, and faculty members who studied these courses for the years from 2010 to 2017. This data has not used to find the strengths, weaknesses, factors, and patterns affecting students' academic behavior during their academic years, such as knowledge of the reasons (excellence, stumbled, failure, and leakage), the level of academic performance of students, the several years spent by students until they graduate.

The field of educational data mining has become a very important tool used at all educational levels, especially in higher education (Manjarres, et. al, 2018). Therefore, in the current study we attempted to conduct an applied study in the field of data mining and statistical inference by devising some factors and patterns to support decision-makers to improve and develop the academic and administrative performance at Majmaah University.

Manuscript received April 5, 2021
Manuscript revised April 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.4.38

According to the current academic advising system at university, we asked the following questions:

Research Question 1: What are the dropout, stumbling, failure, and academically dismissed percentage s?

Research Question 2: What are the best techniques to improve student's academic performance?

The methodological framework adopted in the current study involved four stages: First stage is obtaining academic data for students from the university database. The Second one is preliminary data processing. The third one is choosing the proper method to build the model. Finally, mining data and extract the expected results.

2. Literature reviews

Educational Data Mining using to develop methods to discover unique types of data coming from educational environments (Hegazi & Abugroon 2016; Romero & Ventura 2013;Romero & Ventura 2010). On the other side, Educational Data Mining (EDM) is an emergency discipline that focuses on using Data Mining tools and techniques linked to educational data (Huebner 2013; Jindal and Borah 2013).

Dutt, et. al, (2017) mentioned that Educational Data Mining is concerned with analyzing data generated percentage in an educational setup using disparate percentage systems. Its aims are to develop models to improve the learning experience and institutions' effectiveness. Another study by Smith, et. al, (2019) illustrated an effective academic advising in higher education institutions is paramount to student success, while it is clear that effective advising is positioned to advance the quality of educational programs and their graduates, there is a dearth of supporting evidence, and addressing this through research is a needed priority. There is a positive relationship between students 'ages and their use of academic advising. Some studies advocate that students use academic advisor services in the first-year. On the other hand, female students showed a sharper decline in their participation than male students across their academic life. Therefore, not all students participate equally in academic advising (Roessger, et. al, 2019).

Khobragade, et. al, (2015) attempted to apply Data Mining techniques to predict the failure of students in real-time data from school or graduate students and to detect the failure of students to improve their academic performance and prevent them from dropping out. Different approaches have been applied to solve the high-dimensional problem and use the classification algorithm on earlier and current education information for engineering students to generate the model. This model can be used to detect student academic failure.

Baepler, et. al, (2010) showed the purpose is to conduct academic analyzes and exploratory educational data that quickly produce new possibilities for collecting, analyzing and presenting student data. This study connections the concepts of academic analysis, data extraction in higher education, the investigation of the course management system and suggests how these techniques and the data they produce may be useful to those who practice the process of teaching and learning. The study also said that academic analysis combines selected institutional data, statistical analysis, and predictive modeling to create intelligence that enables students, teachers, or administrators to change academic behavior.

3. Methodology

This study used data warehouse existing in the University to collect and extract inclusive conclusions for provide academic advising on how to improve the GPA at the undergraduate level. In order to build a model to determine the factors that affect students' academic achievement, an analytical descriptive approach was applied to analyse the database, enrolment ratio (GPA_SEC), and cumulative percentage s (GPA). In this work, the methodology is divided into four main phases: data collection, data integrity, data transformation, and data analysis. Fig. 1.below demonstrates the percentages students' academic data model stages.

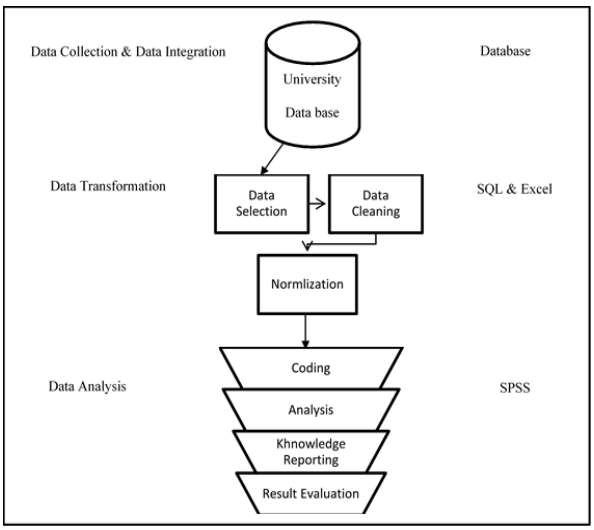

Fig 1. Students' academic data model stages.

Data collection and data integration

In data collection and data integration stage, participants are randomly selected from two colleges at

Majmaah University, college of sciences and humanities in Al-Ghat (CSHG), and college of sciences and humanities in Hotat Sudair (CSHH). Most of the spreadsheets collected were in various formats. The secondary entry percentage tables contains more than 65,000 records. Cumulative GPAs were collected from 16 tables and processed to match the format of other tables.

Data transformation

In data transformation stage, all transformation was handled using SQL. This stage includes three parts, which are data selection, data cleaning, and data normalization (Ahmad, et al, 2015). To answer study's questions, 25 variables were selected. The following steps were taken:

Step 1: GPA_SEC table was linked with the 16 tables, each containing the GPA's for the period from the first semester 2010 to the second semester 2017 that coded (101,102,111,112, up to 172). This step was done for CSHG _ male, CSHG _ female, CSHH _ male and CSHH _ female percentage, these tables named CSG1, CSG2, CSH1, CSH2.
Step2: The tables completed in the first step were linked to the student's status table, which includes the student's status, the duration of the study, department, college, and gender, these new tables named CSG11, CSG22, CSH11, and CSH22.
Step3: by using student ID as a primary key, all variables in step2joined together, 25 parameters were selected to be mined, the parameters are Student _ ID, college, department, student status, gender, location, duration, enrolment semester, GPA_SEC and GPA101 up to GPA172.
All missing values and incomplete data were removed in the data cleaning process.

Data analysis.

In this stage, SPSS statistical package (V. 22) was utilized for data analysis. This stage consists of three phases: coding, analysing, and evaluation of the results. The mean and standard deviation, ANOVA, linear regression and correlation as well as, Chi-square test with a 95% confidence level were used. P<0.05 was considered statistically significant to find the percentage s of excellence, success, repetition, stumbling and leaving study to explore the reasons that led to it and the role of academic guidance to reduce the stumbling, leaving study as well as to give recommendations for increasing excellence.

4. Results and Discussions

The sample size is divided into two parts: 57.6% males and 24.4% females, in which they were classified into 40.8% from CSHG and 59.2% from CSHH. Table 1. below elbow percentages the distribution of data according to gender, college, and specialization.

Table 1. Distribution of data according to Gender, College, and Specialization.

Record Profile	Frequency	Percentage(%)
Gender		
Male	3545	57.6
Female	2613	42.4
Faculty		
CSHG	2511	40.8
CSHH	3645	59.2
Specialist		
IT	608	9.8
ENG	1,289	20.9
MGT	2,412	39.2
I.W.	498	8.1
Iskamic	1,067	17.3
MATH	248	4.0
Chemistry	36	0.6

Table 2. Shows the distribution of students, according to specialization and GPA_SEC. We find that the average percentage of secondary school for registered females was higher than males in all departments except in the Islamic department, where males (M=4.13, SD=0.37) were higher GPA_SEC than females (M=3.96, SD=0.48). There was a significant effect of GPA_SEC on Specialization at the $p<0.05$ level of the seven departments [$F(6, 5797) = 37.233$, $p<0.01$]. There is a statistically significant difference in GPA_SEC between those enrolled in the Islamic department and students who enrolled in IT, ENG, MGT, LW and MATH ($p=0.000$). However, there were no differences between the students that enrolled in IT, ENG, MGT, LW with those enrolled in the MATH department ($p > 0.05$).

The distinction for graduate students was calculated for those whom joined the college for enrolment semester 101 to 141, where excellence approximately 2% of all graduated. Fig. 2 shows the distribution of excellence per graduated and per registered by a semester of enrolment where the highest percentage within enrolment semester was in 131. There was no excellence for graduates registered in 112, while the highest percentage was in semester 131.

Table 2. Distribution of the Mean and Standard deviation of GPA_SEC.

Dept.	Male			Female			Total		
	N	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*
IT	261	4.17	0.41	347	4.28	0.51	608	4.23	0.47
ENG	747	4.24	0.39	542	4.31	0.48	1289	4.27	0.43
MGT	1242	4.23	0.38	1170	4.27	0.43	2412	4.25	0.41
LW	498	4.32	0.40				498	4.32	0.40
Islamic	560	4.16	0.37	507	3.96	0.48	1067	4.06	0.44
MATH	215	4.26	0.45	11	4.44	0.54	248	4.28	0.46
Chemistry				36	4.06	0.50	36	4.06	0.50

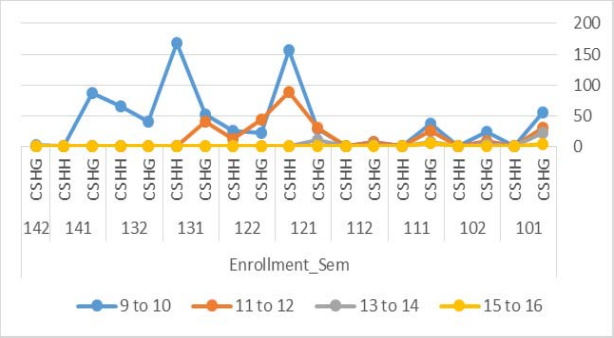

Fig. 2 Distribution of graduated with distinction per enrolment semester**.**

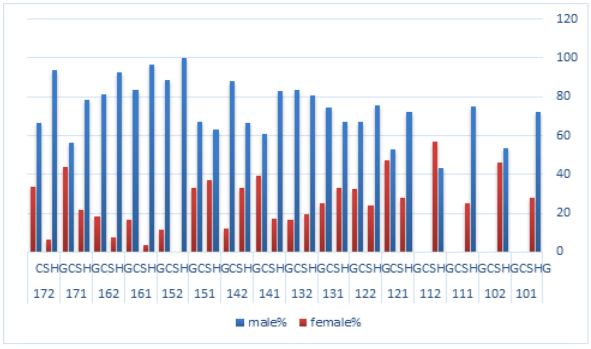

Fig. 3 The percentages by gender 101 -172.

On the other side, students who score GPA less than 2 out of 5 were between the ranges (11.1% to 51.2%), where 51% of them recorded at CSHG in semester 101 and 11.1% recorded at CSHH, CSHG recorded 18.42% in semester 161. Where males scored a higher percentage of failure than females through all semesters with big differences, (see Fig. 3).

Fig. 4 and Fig. 5 show the number of graduates, according to enrolment semester and duration of study. The results show that the percentage of graduated students in less than eight semesters was 20% in the period from 101 to 141, while 97.8% of them were from CSHH, 37.5% were graduated after complete exactly eight semesters. In addition, the rest (42.5%) graduated after completing nine to twelve semesters. It indicates that more than 62.5% of graduated students failed to follow the study plans.

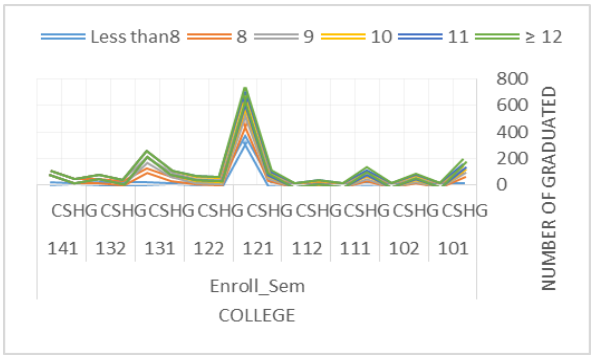

Fig. 4 Graduate Students per period.

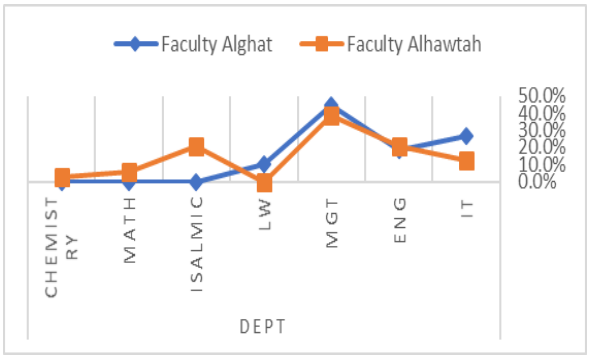

Fig. 5 Graduate percent according to the department during enrolment from 101 to 141.

Fig. 6 shows that dismissal percentage was 5.2% through the period of the study. The distribution of dismissal percentage was 57.8% for CSHG and 42.2% for CSHH. Males registered the highest percentage with 73.1%, while CSHG _ male represents the highest percentage with 49.1% of the total compared with CSHH-male that scored 24.1. The lowest dismissal percentage was 8.8% for CSHG-female. The majority of 27.8% dismissed after three semesters followed by 24.7% dismissed after completion of two semesters. There was a weak negative correlation between gender and GAP-SEC ofdismissed r (319) =-0.45, P<0.01) and observed a weak positive relationship between Dept. & college, and gender & college r (319) =-0.31, P<0.01), r (319) =-0.4, P<0.01) respectively.

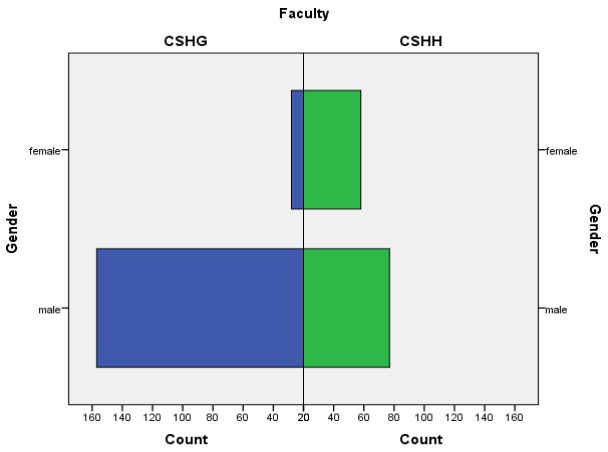

Fig. 6 Dismissed distribution according to colleges.

We found that 22.9% of students were either dropped, withdrawn, leaved, apologized, or deferred (Fig. 7). Dropout distribution for students who left was (5.9%), withdrawn was 16%, deferred was 0.2%, and apologized who did not return to study was 0.8%. Withdrawal represents the main category for dropped out students in which it represents 16.7% of participants, 986 students. The ones who withdrawn after the completion of one semester was 40.7%, where those who withdrawn after the completion of two semesters was 21.1% for the period between 101 to 172 semesters, as shown in Figure 8. The main effect of GPA was found a significant difference in GPA for males (M=1.177, SD=1.08) females (M=1.71, SD=1.01): t (117) =-2.619, p<0.01.

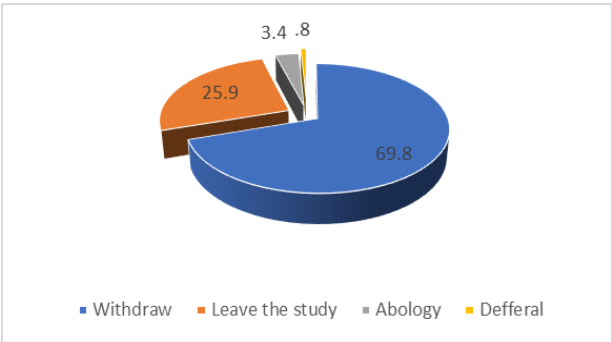

Fig. 7 Dropout percentage according to the type of during 2010 – 2017.

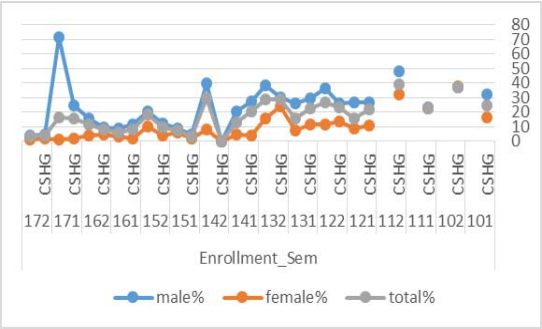

Fig. 8 Withdraw within gender and college from 101 to 172.

Students who were graduated or who were stayed at college for more than 8 semesters around 1116, representing 32% of the students who enrolled for 101 to 141. The students who studied 9 to 10 semesters were 69.5%, who studied 11 to 12 semesters was 25.4%, who studied 13 to 14 was 4% and only 0.8% who studied 15 to 16 (as illustrated percentage in Figure 9). The mean GPA for the first semester lies between (M=1.82, SD=0.97) and (M=3.11, SD=1.07). The multiple linear regression is used to predict the participants GPA of enrolled students up to 141 semester. Students GPA in earlier semesters, their college, department, and gender are fed to backward method. Three variables were remaining significant in the model. A significant regression equation was formed F (3,77) =27.539, p<0.01 with R^2=0.518, participant predicted GPA is equal to 0.033+0.180x1+0.227x2+0.580xn-1, where x1, x2 are GPA's of the first and second semesters and xn-1 is the before last semester.

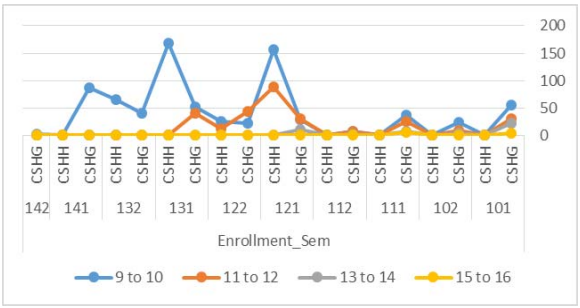

Fig. 9 Stumbling Students by the number of completion semesters.

Conclusions

This work showed the percentage of Excellency, stumbling, failure, dismiss and dropout. The percentage of excellence was 2% for all graduates. The results showed that the percentage of failure among males was higher than that of females during all semesters. Graduated students in less than eight semesters scored 20%, where 97.8% of them were from CSHH, 37.5% were graduated after completing exactly eight semesters, where the rest (42.5%) graduated after completing nine to twelve semesters, which indicates that more than 62.5% of graduated students failed to follow the study plans. A percentage of 5.2% of the overall total was dismissed with the weak positive relationship between Dept. & college, and gender & college r (319) =-0.31, P<0.01, r (319) =0.4, P=0.01 respectively, the majority dismissed after completing two to three semesters. The dropout percentage represented about 22.9% of the total number of students, representing those who were withdrawn, leave, apologize and deferral those who did not return to study. We recommended to decision-makers for raises the level of academic advisors by including programs to follow up on stumbling and failure by the end of each semester to learn the reasons and address them to reduce the percentage of dismissed and dropout. The study was conducted at two colleges from Majmaah University with a convenience sample. Future studies should be conducted on samples with different colleges in the University to see if similar findings will be held. Also, should address the construction of an organized database for all students, courses and programs, as well as the development of a program to raise the level of underprivileged students supervised by academic advisers.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

The authors are equally contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Acknowledgment

The authors would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project Number No. (R-2021-100).

References

[1] Abu-Oda, G. S., & A. M. El-Halees, Data Mining in higher education: University student dropout case study, International Journal of Data Mining & Knowledge Management Process 5(1) (2015), 15-27.

[2] Agnihotri, A., & B. Mishra, *Application of Data Mining Techniques in Higher Education (With Special Reference to Improving the Quality)*, (IJCSIT) International Journal of Computer Science and Information Technologies 6(6) (2015), 4855-4859.

[3] Ahmad, F., Ismail, N., & A. A. Aziz, *The prediction of students' academic performance using classification Data Mining techniques*, Applied Mathematical Sciences 9(129) (2015), 6415-6426.

[4] Ali, S. M., & M. R. Tuteja, *Data Mining Techniques*, International Journal of Computer Science and Mobile Computing 3(4) (2014), 879-883.
[5] Aslam, S., & I. Ashraf, *Data Mining algorithms and their applications in education Data Mining*, Int. J 2(7) (2014), 50-56.
[6] Dutt, A., M. A. Ismail & T. Herawan, *A systematic review on educational Data Mining*, IEEE Access 5 (2017), 15991-16005.
[7] Guleria, P., & M. Sood, *Data Mining in Education: A review on the knowledge discovery perspective*, International Journal of Data Mining & Knowledge Management Process 4(5) (2014) 47-60.
[8] Hegazi, M. O., a& M. A. Abugroon, *The state of the art on educational Data Mining in higher education*, International Journal of Computer Trends and Technology 31(1) (2016), 46-56.
[9] Huebner, R. A, *A Survey of Educational Data-Mining Research, Research in higher education*, journal 19 (2013), 1-13.
[10] Jain, S., R., Raghuvanshi, & M. Ilyas, *A Survey Paper on Overview of Basic Data Mining Tasks*, International Journal of Innovations & Advancement in Computer Science 6(9) (2017), 246-256.
[11] Jindal, R., M. D. Borah, *A survey on educational Data Mining and research trends*, International Journal of Database Management Systems 5(3) (2013), 53-73.
[12] Khobragade, L. P., & P. Mahadik, *Students' academic failure prediction using data mining*, International Journal of Advanced Research in Computer and Communication Engineering 4(11) (2015), 290-298.
[13] Manjarres, A. V., L. G. M. Sandoval & M. S. Suárez, *Data Mining techniques applied in educational environments: Literature Review*, Digital Education Review 33 (2018), 235-266.
[14] Roessger, K. M., Eisentrout, K., & Hevel, M. S, *Age and academic advising in community colleges: Examining the assumption of self-directed learning, Community College*, Journal of Research and Practice 43(6) (2019), 441-454.
[15] Romero, C., & S. Ventura, *Data mining in education*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3(1) (2013), 12-27.
[16] Romero, C., & Ventura, *Educational data mining: a review of the state of the art*, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40(6) (2010), 601-618.
[17] Rupali, G. G, *Data Mining: Techniques, Applications and Issues*, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) 2(2) (2013), 220-223.
[18] Shruthi, P., & B. P. Chaitra, *Student Performance Prediction in Education Sector Using Data Mining*, International Journal of Advanced Research in Computer Science and Software Engineering 6(3) (2016), 212-218.
[19] Baepler, P., & Murdoch, C. J, *Academic analytics and data mining in higher education*, International Journal for the Scholarship of Teaching and Learning 4 (2010) ,1-9.
[20] Smith, D. T., Broman, T., Rucker, M., Sende, C., & Banner, S, *Advising in kinesiology: Challenges and opportunities*, Kinesiology Review 8(4) (2019), 323-328.
[21] Suhirman, S., T. Herawan, H. Chiroma, & J. M. Zain, *Data Mining for education decision support: a review*, International Journal of Emerging Technologies in Learning (iJET) 9(6) (2014), 4-19.
[22] Thakar, P, *Performance analysis and prediction in educational Data Mining: A research travelogue*, arXiv preprint arXiv 110(15) (2015), 60-68.