International Journal of Advanced Science and Research
ISSN: 2455-4227; Impact Factor: RJIF 5.12
Received: 27-01-2020; Accepted: 21-02-2020; Published: 28-02-2020
www.allsciencejournal.com
Volume 5; Issue 1; 2020; Page No. 36-40

Outlier and extreme values detection to evaluate students performance using data mining techniques

Prashant G Tandale[1], Anil T Gaikwad[2]
[1] Assistant Professor, Bharati Vidyapeeth Institute of Management, Kolhapur, Maharashtra, India
[2] Associate Professor, Bharati Vidyapeeth Institute of Management, Kolhapur, Maharashtra, India

Abstract
Educational data mining is an emerging area that focuses on applying data mining tools and techniques to educationally related data. Data mining is considered as the most suited technology appropriate in giving additional insight for the different stake holders of an educational institution by discovering the hidden patterns, associations, and anomalies from educational data and acting as an active automated assistant in helping them for making better decisions on their educational activities. This paper tries to take review of work done so far in this area of educational data mining and one of the applications of data mining called "Outlier and Extreme Value Detection".

Keywords: educational data mining (EDM), outlier, extreme values, covid-19 pandemics

1. Introduction
Education is the key to the prosperity of any nation. India is one of the fastest growing nations in the world with the largest youth of population [19]. Hence, in order to build a skilled workforce education becomes necessary. Students are opting for the fields such as engineering, science, and technology. Unfortunately due to lack of quality education at primary level (The economist, 2008), socio-economic, psychological and other diverse factors, students' failure rates are high and performance is low. Hence to improve the quality of engineering graduates, such cases of failure and poor performance must be monitored proactively. Initially, the applications of data mining were restricted to the business domain but now it is extended to education and is known as Educational Data Mining (EDM). EDM deals with the application of data mining tools and techniques to inspect the data at educational institutions for deriving knowledge [2]. The data mining techniques can help the institutes in extracting patterns like students having similar characteristics, Association of students' attitude with performance, what factors will attract meritorious students and so on. The past several decades have witnessed a rapid growth in the use of data and knowledge mining as a means by which academic institutions extract useful hidden information in the student result repositories in order to improve students' learning processes [14].

2. Review of Literature
A literature review on educational data mining topics such as student retention and attrition, personal recommender systems within education, and how data mining can be used to analyze course management system data.
According to Paulraj and Ponniah, the main benefits of data mining to educational institutes are – It provides an integrated and total view of an institute. It makes the institute's current and historical information easily available for the decision making [13].
Agarwal, Pandey & Tiwari suggested that student's placement is based on his performance in qualifying examination and test marks. Since placement is one of the most important parameters for quality of education, it is immensely necessary that students' performance must be improved which is our area of focus throughout the paper [1].
C. Romero and S. Ventura carried out a survey for education field. They have described the types of users, types of educational environments and the data they provide. Also they have explained in their work the common tasks in the educational environment that have been resolved through data mining [16].
Shaeela Ayesha & others discusses data mining technique named k-means clustering is applied to analyze student's learning behavior. Here K-means clustering method is used to discover knowledge that come from educational environment [17].
Kamal, Chowdhury & Nimmy had used enrollment data to predict the dropout of Information Systems students studying in the department of Computer Science and Engineering (CSE), University of Chittagong. They had used Bayes theorem based on the knowledge base to predict the dropout. [9].
Robertas analyzed student academic results for informatics course improvement, rank course topics following their importance for final course marks based on the strength of the association rules and proposed which course specific course topic should be improved to achieve higher student learning effectiveness and progress [14].
W.M.R. Tissera & others presents a real-world experiment conducted in an ICT educational institute in Sri Lanka. A series of data mining tasks are applied to find relationships between subjects in the undergraduate syllabi. [18].
Hongjie Sun conducts a research on student learning result based on data mining. It is aimed at putting forward a rule-discovery approach suitable for the student learning result evaluation and applying it into practice so as to improve learning evaluation skills and finally better serve learning practicing [8].

S. Anupama Kumar and Dr. Vijayalakshmi M.N applied decision tree algorithm on student's internal assessment data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to fail or pass [11].

Monika Goyal used different types of rule –based systems and have been applied to predict student's performance (mark prediction) in an e learning environment (using fuzzy association rules). [5].

According to Jaiwei Han, data mining is an interdisciplinary field of astronomy, business, computer science, economics and others to discover new patterns from large data sets. The actual data mining task is to analyze large quantities of data in order to extract previously unknown patterns such as groups of data records (cluster analysis), unusual records(anomaly detection) and dependencies (association rule mining) [6].

3. Educational Data Mining

Data mining is a series of data analysis techniques applied to extract hidden knowledge from server log data [15]. by performing two major tasks: Pattern discovery and predictive modeling [12]. Educational data mining (EDM) is a field which adopts data mining algorithms to solve educational issues. Romero & Ventura reviewed 306 EDM articles from 1993 to 2009 and proposed desired EDM objectives based on the roles of users [16]. For the purpose of this study, which is designed to inform administrators, the list is limited to objectives for administrators:

a. Enhance the decision processes in higher learning institutions
b. Streamline efficiency in the decision making process
c. Achieve specific objectives
d. Suggest certain courses that might be valuable for each class of learners
e. Find the most course effective way of improving retention and grades
f. Select the most qualified applicants for graduation
g. Help to admit students who will do well in higher education settings

Outlier Analysis

Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains' [3]. Outlier detection has been a widely researched problem and finds immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, intrusion detection for cyber security, adverse drug reaction, fault detection in safety critical systems, military surveillance for enemy activities and many other areas [7, 10].

3. Applications of Data Mining Techniques in Higher Education

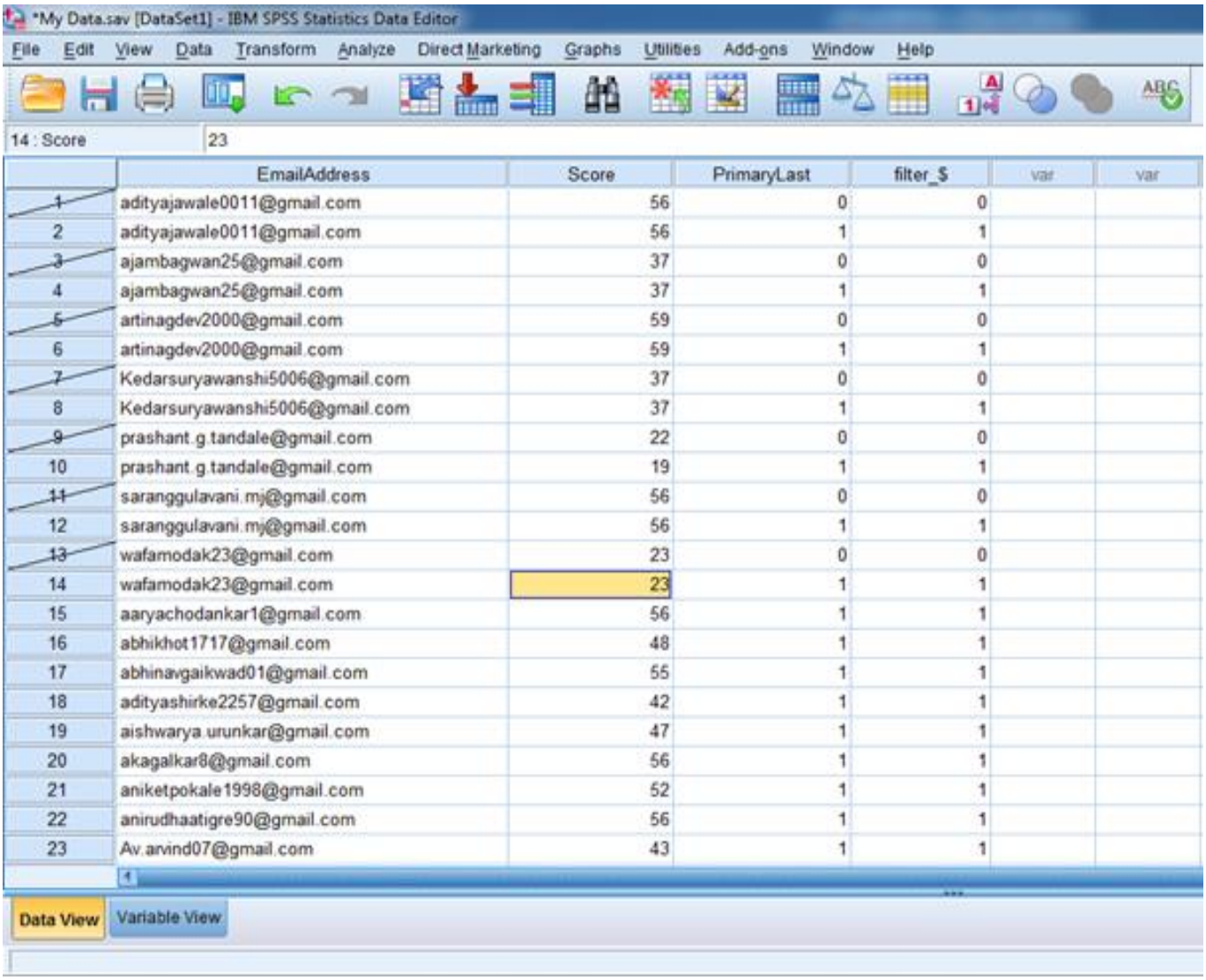
A. Problem Description and Solution

In continuation of UGC's letter dated 19th March, 2020 whereby all the ongoing examinations and evaluation work were required to be postponed till 31.03.2020 in the light of Novel Corona virus (COVID-19) outbreak. Faculty members under all Universities are permitted and advised to work from home [20].

During the same period our University internal examination were scheduled. The challenge in front of the faculty members was how to conduct the examinations on time as well as how to avoid the academic loss of the students. To overcome this situation, the researcher came up of the idea of online Google Quiz using Google forms. So the online MCQ type test was conducted by the researcher while informing the students 3 days in advance. The stipulated time was selected and test was started at that particular time and ended on time by not accepting the responses from the respondents. Overall 72 responses are registered. By using the statistical tools and techniques noise detection is done and duplicate or repeated responses are eliminated. After this filtering 64 responses were remaining. The maximum marks for the test was 60. Now the real challenge was to find the outliers and extreme values the obtained data so that we can the necessary academic decisions.

B. Detection of Outliers and Extreme Values in an Educational Data

For this purpose Online Test Data of third year UG Students of BDVU, Institute of Management, Kolhapur is collected using Google Quiz. The collected data is analyzed using SPSS software. We used filters to remove the noise from the data like duplicate responses. Then we applied the following procedure to find out the Outliers and the extreme values in the test data. The screenshot of the data analysis is as shown in tables and graphs below.



	EmailAddress	Score	PrimaryLast	filter_$	var	var
1	adityajawale0011@gmail.com	56	0	0		
2	adityajawale0011@gmail.com	56	1	1		
3	ajambagwan25@gmail.com	37	0	0		
4	ajambagwan25@gmail.com	37	1	1		
5	artinagdev2000@gmail.com	59	0	0		
6	artinagdev2000@gmail.com	59	1	1		
7	Kedarsuryawanshi5006@gmail.com	37	0	0		
8	Kedarsuryawanshi5006@gmail.com	37	1	1		
9	prashant.g.tandale@gmail.com	22	0	0		
10	prashant.g.tandale@gmail.com	19	1	1		
11	saranggulavani.mj@gmail.com	56	0	0		
12	saranggulavani.mj@gmail.com	56	1	1		
13	wafamodak23@gmail.com	23	0	0		
14	wafamodak23@gmail.com	23	1	1		
15	aaryachodankar1@gmail.com	56	1	1		
16	abhikhot1717@gmail.com	48	1	1		
17	abhinavgaikwad01@gmail.com	55	1	1		
18	adityashirke2257@gmail.com	42	1	1		
19	aishwarya.urunkar@gmail.com	47	1	1		
20	akagalkar8@gmail.com	56	1	1		
21	aniketpokale1998@gmail.com	52	1	1		
22	anirudhaatigre90@gmail.com	56	1	1		
23	Av.arvind07@gmail.com	43	1	1		

Fig 1: Finding and Removing Duplicate Data

Table 1: Statistical Data Analysis

Mean	Median	Variance	Std. Deviation	Inter-quartile Range	Skewness	Kurtosis
49.81	54.00	147.01	12.12	10	-2.05	4.22

Table 2: Extreme Values and Outliers

Score		Student Roll Number	Marks Obtained
Highest	1	35	60
	2	36	60
	3	39	60
	4	41	60
	5	42	60
Lowest	1	69	7
	2	58	7
	3	10	19
	4	32	21
	5	14	23

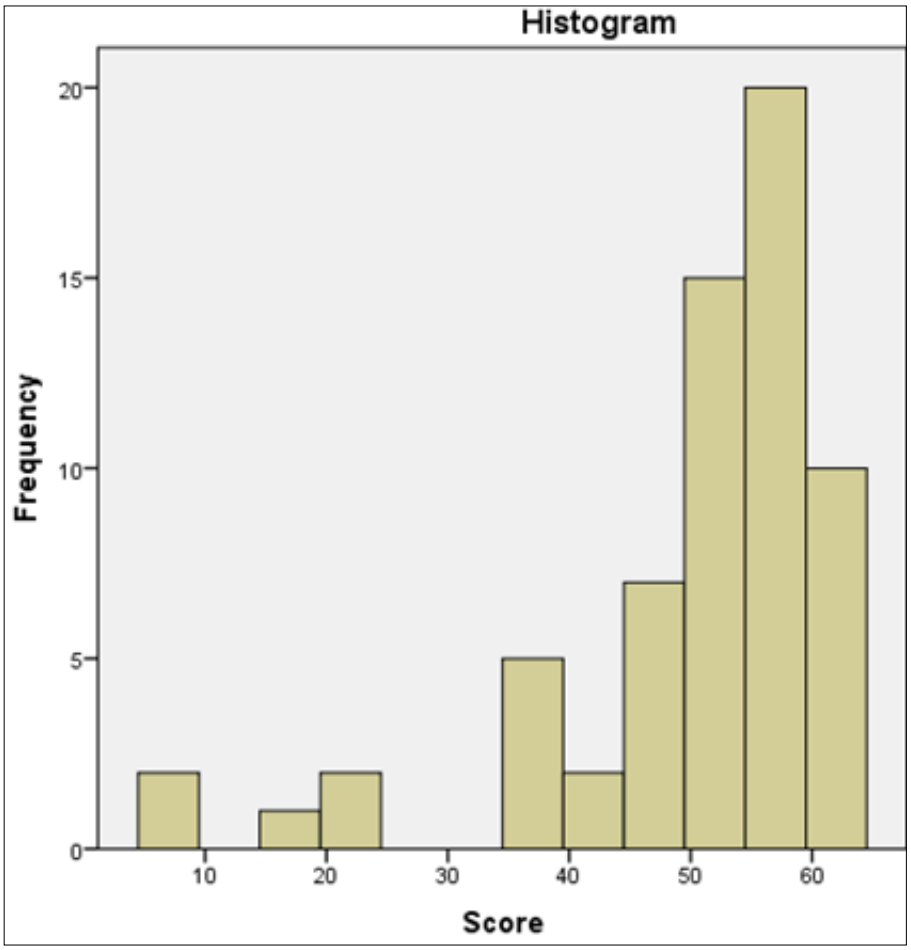


Fig 2: Histogram showing score frequency of the Test Data

C. Data Analysis and Findings

From Fig. 1 we observed that there are some duplicate records. We used filters from SPSS to remove the duplicates. Table 1 shows basic statistical information obtained after applying statistical test to the students data. In Table 2 we obtained extreme values and outliers. From all above figures and table we can interpret the following results.

a. Average score of the test is 49.81 out of 60, means that average subject score of the all students is 83.01%. This indicates overall performance of the students is excellent.
b. Median is 54.0, means the middle observation is 54.0, from the graph we can see that most of the score is ranging between 30 and 60.
c. The graph or histogram indicates that the entire data is right skewed that is -2.05 since mean is less than the median.
d. Standard deviation is 12.12, means there is very less differences between the observations, that is all the observations are very closed to each other. Means students scores are much closed to each other.
e. Now most important part of this research is Table 2, which shows extreme values and outliers. Here the extreme values are 7 as left extreme and 60 as the right extreme. This observation forces us to think on this values regarding is there is any suspicious activity i.e. outliers, needs to be discussed in a academic perspective.
f. Roll numbers of the students those are having lowest score are considered for further study like their attendance, previous examinations performance, etc. Mean we can trace the possible reasons for their poor performance and accordingly we could plan the academic activates for the slow learners.
g. Similarly Roll numbers of the students those are having Highest score are considered for further study like their attendance, previous examinations performance, etc. Mean we can trace the possible reasons for their Extremely High performance and accordingly we could plan the academic activates for the fast learners or bright students. Here there may be possibility of organized malpractice activity undertaken by the students, so the teacher may discuss personally with these students regarding their performance in the test.
h. It is also comes to know from that test analysis that there may be possibility that some students facing some problem regarding availability to internet strength, bandwidth, electrical connection as well as some technical problem in the devices they are using like mobile phones, laptops, etc.
i. Also we observed that Roll numbers of the students with highest score are in very closed range from 35 to 42, which will may raise the flag of suspicious activity. This way we can evaluate the performance of the students in the test carried out during the COVID -19 lockdown period.

4. Conclusion

Let us summarize the study we have done so far. We used various tool and techniques of data mining to find out extreme values and outliers in the given data, which are

used to evaluate the academic performance of the students. It is very necessary to use technology to do conduct the tests and assessment. Because of this method we can get the academic test results very fast. Students are showing interest in such type of test as compare to offline or written class tests, which indicated by the average score, standard deviation of the test. On the contrary some student's performance is extremely high and some students performance is very poor. The designers of the test must maintain the equal balance of the difficulty level of questions while setting the MCQ type question paper. Similarly it is very necessary to study how much comfortable the students are in case of online tests conducted during COVID-19 lockdown period. Findings of this paper may be considered as some guidelines to the educational institutions regarding conducting online tests for students in future in such pandemics situation.

5. References

1. Agarwal PT. Data Mining in Education: Data Classification and Decision Tree Approach. Journal of e-Education, e-Business, e-Management and e-Learning, 2012, 140-144.
2. Al-razgan Ak. Educational Data Mining: A systematic review of the published leterature 2006-20013. Internatinal Conference on Advanced Data and Information Engineering, 2013, 711-719.
3. Chandola VBA. Anomaly Detection: A Survey. ACM Computing Surveys. 2009; 41(3):15.
4. Creaking groaning Infrastructure is india's biggest handicap. The Economist, 2008.
5. Goyal M. Applications of Data Mining in Higher Education. International Journal of Computer Science. 2012; 9(2):113-120.
6. Han J. Data Mining – concepts and Techniques. Elsevier Publication, 2012.
7. Hodge VJA. A survey of outlier detection methodologies. Artificial Intelligence Review. 2004; 22(2):85-126.
8. Hongjie S. Research on Student Learning Result System based on Data Mining. International Journal of Computer Science and Network Security, 2010, 4(10).
9. Kamal C. New Dropout Prediction for Intelligent System. International Journal of Computer Applications. 2012; 42(16):26-31.
10. Koh YSPR. Rare association rule mining via transaction clustering. Proceedings of the 7th Australasian Data Mining Conference. Australian Computer Society, Inc. 2008; 87:87-94.
11. Kumar SA. Mining of student academic evaluation records in higher education. International Conference on Recent Advances in Computing and Software Systems (RACSS) IEEE, 2012, 67-70.
12. Panov. Towards an ontology of data mining investigations. In S. L. Panov P., Lecture Notes in Artificial Intelligence, 2009, 257-271.
13. Paulraj. Data Warehousing Fundamentals: A Comprehensive Guide to IT Professionals. John Wiely & Sons, 2001.
14. Robertas D. Analysis of Academic Results for Informatics Course Improvement using Association Rule Mining. Springer US, 2009.
15. Roiger RJG. Data mining: a tutorial-based primer. Boston, MA: Addison Wesley, 2003.
16. Romero C. Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, 2010, 6(40).
17. Shaeela A. Data Mining Model for Higher Education System. European Journal of Scientific Research. 2010; 1(43):24-29.
18. Tissera W. Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining. International Conference on Information Acquisition. IEEE, 2006.
19. Unfa. The Power of 18 Billion Adolescents, Youth and Transformation of the Future. EECA, 2014.
20. UGC's. letter dated 19th on subject " Precautions to be taken in the light of Novel Corona virus (COVID -19), 2020.
21. Hung JL, Hsu YC, Rice K. Integrating Data Mining in Program Evaluation of K-12 Online Education. Journal of Educational Technology & Society. 2012; 15(3):27-41.
22.