CrossMark

Technical Communication

Data Mining and Knowledge Discovery in Big Data for Decision Making in Higher Education

Hanna Mohammad Said[1]* and Abdel-Badeeh M. Salem[2]

[1]Department of Business Management and Information System, College of Science and Humanities at Alghat, Majmaah University, Al-Majmaah 11952, Saudi Arabia

[2]IT, Cairo, Egypt, Ahram Canadian University, Cairo, Egypt Faculty of Computer and Information Sciences- Ain Shams University Cairo, Egypt

ABSTRACT

Artificial intelligence and data mining plays a fundamental role in improving the intelligence of education through special standards for improving teaching quality, better learning experience, predictive teaching, assessment method, effective decision-making, and improved data analysis. BD (Big Data) are also used to assess, detect, and anticipate decision-making, failure risk, and consequences to improve decision-making and maintain high-quality standards. According to the findings of this study, certain universities and governments have adopted BD to help students transition from traditional to smart digital education. Many obstacles remain in the way of complete adoption, including security, privacy, ethics, a scarcity of qualified specialists, data processing, storage, and interoperability. Learning today is getting smarter, thanks to the rapid development of the use of data and knowledge for big data analysis. Besides delivering real-world knowledge discovery applications, specialized data mining methodologies, and obstacles have real-world applications. Therefore, this article aims to explain the current concept of an intelligent learning environment in higher education. It explores the main criteria, and presents evaluation methods through the use of the proposed model.

KEY WORDS: ARTIFICIAL INTELLIGENCE, DATA MINING, KNOWLEDGE DISCOVERY, BIG DATA, AND E-LEARNING.

INTRODUCTION

Data-driven decision-making is rapidly evolving in education, it has quickly evolved into the far more sophisticated idea of big data, which is based on software techniques known as predictive analytics. Big data and analytics for educational applications are still in their infancy and will take a few years to mature, but they are already making an impact and should not be overlooked. While big data and analytics aren't a silver bullet for all of the problems and decisions that higher education administrators confront, they can be integrated into many educational periodicals' embedded solutions. The goal of this research was to expand Saudi higher education's use of big data and analytics. It looks at the nature of these notions, present basic definitions, evaluate potential applications, and, last but not least, highlight issues with their implementation and expansion (Alsheikh2019; Aljahdaliet al., 2020; Alkhalil, 2021).

Data mining is a method for identifying previously undiscovered associations in large amounts of data. Data mining and this technique is used as a useful approach for many areas of application in educational settings, particularly in higher education, but has not been generally accepted by users of the educational industry in general and the database community in particular. The nature of data mining jobs is primarily responsible for this. Data mining is a challenging and time-consuming task that requires a lot of practice to succeed, (Kaiser et al 2013).

Data mining usually involves selecting techniques, creating models, and tuning parameters to support analytical activities. Data mining technology helps companies make use of their existing data more effectively and gain insightful information that gives them a competitive advantage. Data mining is the process of extracting hidden predictive information from massive databases. It's a strong tool that may assist higher education institutions to focus on the most significant data in their data repositories while making decisions. Data mining techniques enable all higher learning institutions to make proactive, knowledge-based decisions by predicting

2013

trends and behaviors. Automated and prospective analyzes of data mining go beyond analyzes of past events presented by potential tools used in decision support systems, (Kaiser et al., 2013; Gandomi et al., 2015).

Data mining techniques also enable to quickly answer problems that previously took a long time. They create databases to search for hidden patterns and prognostic information that experts may overlook since it is outside of their expectations. The proposed use of association rules, simple decision tree, and Logistic Regression Algorithms as means for data mining is a useful technique for many risks that can face higher education systems. Existing studies related to intelligence analysis have focused on news analysis or information analysis forums, but unfortunately, few studies use data mining and knowledge discovery techniques from big information in education(Picciano2012; Muktharet al., 2017; Baiget al., 2020). This paper consists of 3 sections: The first section includes a brief overview of data mining and knowledge discovery, and the second section includes the data mining methodology, decision tree model, and logistic regression algorithms. Section 3 includes the conclusion and future work.

Data Mining and Knowledge Discovery in "big data: There are many applications for data mining and knowledge discovery. Many conferences on Artificial Intelligence, Intelligent Systems, Databases, and Statistics have discussed data mining and knowledge discovery. The process of detecting validated, novel, and plausible patterns is referred to as knowledge discovery. Data mining, or discovering information from big data, is the process of applying the discovery process to large databases or data sets. Creating the target dataset, building an understanding of the application domain, etc. are all stages in the discovery process.; finding useful features for data representation; data cleaning and preprocessing; It uses data mining to look for patterns of interest; Interpretation and standardization of the detected patterns ((Muktharet al., 2017; Aljahdaliet al., 2020; Alkhalil, 2021).

From an obvious point of view, many algorithms seem to be completely different, but from the point of view of reality, they have common components that make them similar. To decode and understand data mining and model extrapolation contribute to decision making, it is necessary to delve into the level of these components. This, in turn, will lead to an understanding of the overall contribution and good application of the KDD process as shown in Figure 1.

Big Data Analytics' Importance in Higher Education: Big Data Analytics (BDA) is the first phase in the knowledge discovery process for Higher Education, and it involves the algorithms that are used to extract potentially significant patterns, connections, trends, sequences, and dependencies from data, as shown in Figure 2.

1. Determination: Getting to know the application domain:
2. Preparation: Cleaning and preprocessing of data:
3. Data reduction and projection (transformation):
4. Data Mining: Applying Data Mining Functions and

Algorithms: Rules of association, classifying,
5. and grouping
6. Validation and Verification of Patterns: Validate and verify identified patterns:
7. Applying everything you've discovered:

A. Data cleaning (a) (to remove noise and inconsistent data)
B. Data fusion (where multiple data sources may be combined)
C. Data selection (data from the database that is relevant to the analytical activity)
D. Data transformation (conversion or aggregation of data into acceptable forms for a certain purpose).
E. Data mining is the process of finding and extracting meaningful patterns from massive volumes of data(Data visualization).
G. Results verification and assessment; formulating conclusions

**Figure 1: The Process of Knowledge Discovery (Aljahdaliet al., 2020).**
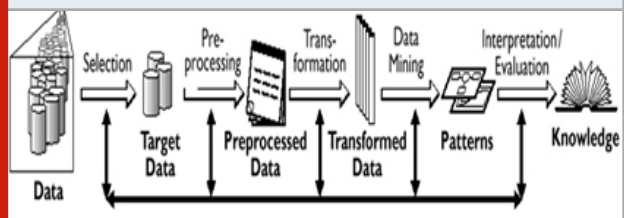
Figure2: Big Data Analytics' Importance in Higher Education (Boulilaet al., 2018).
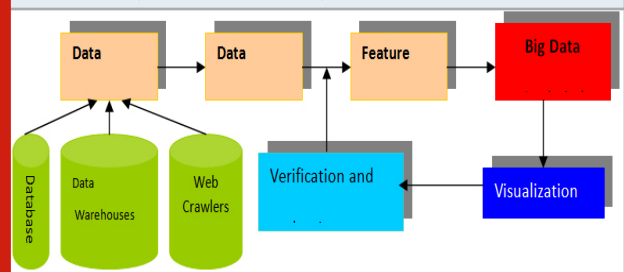
This section describes the main aims of data mining, particularly in the higher education sector, as well as the approaches utilized to attain these goals and the data mining algorithms employed in these procedures. Are thefirst, third, fourth, and fifth numbers. Knowledge discovery goals can be described in terms of the system's intended usage. Two fundamental aims can be distinguished(Boulilaet al., 2018; Vermaet al., 2018; Alkhalilet al., 2021). 1) Validation, 2) Investigation.

Challenges: Big data information systems, especially electronic higher education systems, are vulnerable to a variety of threats. Since information is exposed to a variety of risks and attacks through the collection, processing, and retrieval stages - whether reading, printing, or downloading - as well as the transmission, exchange, and storage stage, these risks, and attacks differ according to these. Operations,

each stage has its risks and means of protection (Aljahdaliet al.,2020; Aseeriet al., 2020).

Figure 3: Fields of DM Combination
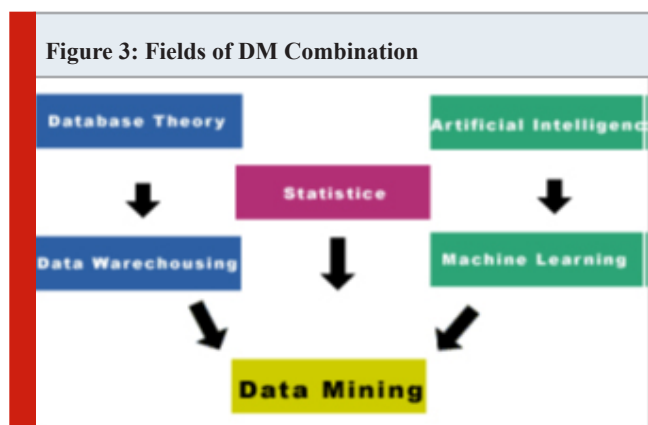
System penetration: This occurs when an unauthorized individual gains access to a computer system and engages in unlawful actions such as changing application software, stealing private data, damaging files, software, or the system, or simply for illicit purposes. Or by exploiting system flaws, such as circumventing control and protection mechanisms, or by using knowledge gathered by the hacker from material or moral sources. These are the most notable threats and assaults (Menget al., 2014).Cultivation of flaws: Typically, this threat arises as a result of an illegal entry or a legitimate user beyond the scope of the authorization provided to him., so that the individual creates an opening for eventual penetration (Segooaet al., 2018).

Usage of the right of authorization: This occurs when a person who is authorized to use the system for one purpose uses it for another without first getting permission. External threats are mitigated by exploiting a flaw in the system to gain access to it through a valid method or a legitimate element and then engaging in criminal actions (Ndaet al., 2019). Monitoring communications: By merely monitoring communications from one of the victims' computers, the offender can get sensitive information, which is typically information that helps further infiltration of the system, (Baiget al., 2020).

Concept of Data Mining: Data mining is a logical technique for searching through enormous volumes of data to identify essential information, that shown in figure 4. The purpose of this method is to discover previously undiscovered patterns. You may utilize these patterns to address a variety of issues after you've discovered them. Data mining (knowledge discovery) is the process of uncovering data from many angles and presenting it in a form of value that can be used to enhance its value, reduce costs, or both (Al-Medlej, 1997; Segooaet al., 2018).

Data mining software is one of the many analytical methods available for data analysis, especially big data. Higher education groups enable the study of data from a variety of viewpoints, their classification, and the description of the links discovered in them to make the appropriate decision. Data mining is the process of identifying patterns or links between hundreds of variables in large databases. Data

mining is a strong tool since it may supply businesses with pertinent information that can be used to their advantage. When businesses have the appropriate knowledge, all they have to do is use it in the right way to get the desired outcome (Janssenet al., 2017).

Nowadays, obtaining information is pretty simple. However, obtaining relevant information that might assist you in achieving the desired objective is not always simple. This is when data mining comes into play. This is where data mining becomes an important technique to master. It has the capability of predicting certain system behavior (Menget al., 2014).

1. Model Validation and Evaluation in Statistics
2. Pattern recognition and machine learning: Symbolic neural network learning
3. Data warehousing and databases
4. The Algorithm (online analytical processing)
5. Knowledge representation in artificial intelligence
6. Report production using data visualization
7. Recognizing Patterns

Figure 4: Knowledge Discovery and Data Mining (Mukthar,2017)
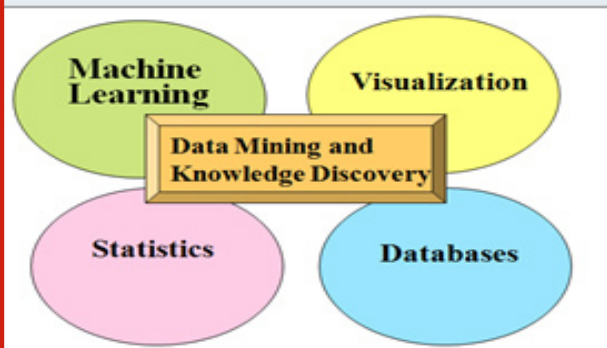
Data mining techniques may be divided into two categories: tasks and methods (Hashim et al., 2010).

Methodology: The proposed use of association rules, simple decision trees, and Logistic Regression Algorithms as data mining tools is a useful technical method for assessing data security, classifying and analyzing data, and ensuring that it is not transmitted to provide strategic information for the various services provided, as well as finding important management points. Effective educational services, what kind of data should be utilized, what type of data was communicated properly, what are the conditions or needs of the user in the data, knowledge organization and arrangement in terms of time, priority, and so on.

Classification of activities, or in terms of its importance or the moment of its discovery Has there been a breach of information, the underlying structure for sharing knowledge, which leads to the addition to the network based on the numerous efforts made to the mechanism of this process to explore, analyze data and test its integrity and within the framework of prediction, based on the criteria used, which help in choosing between many alternatives? The challenges

and risks that higher education systems face. Algorithms for mining spatial, textual, and other complex data.

1. Methods of incremental discovery and re-use of previously acquired information.
2. Methods of discovery are combined.
3. Data mining data structures and query evaluation techniques
4. Data mining approaches that are parallel and distributed.
5. Problems and obstacles in working with large or small data sets
6. Statistics, databases, optimization, and information processing are fundamental topics as they apply to the difficulties of extracting patterns and models from data.

The practical fundamental goals of data mining may be summed into two basic goals: "prediction" and "description."

1. Data mining prediction entails using certain database variables to forecast future values of other variables of interest.
2. The search for human-interpretable patterns that can characterize the data is the focus of the description process. To put it another way, descriptive modeling (clustering) is the grouping of related data into a single group.

Verification and testing of availability Using rules of association: The use of the term"availability" in communications as a factor to measure and describe the degree to which cyberspace remains functional. In most cases, it indicates a fraction, such as 9998 for the available simple. A, and in the case of subtracting from it the downtime (the time taken), (given the percentage of the expected work value & performance).

$$A = \frac{E[UPtim]}{E[Downtim]} + \frac{E[UPtim]}{E[UPtim]}$$

If you define a function, it will be x (t). At Tim t, Status 1 = sys Function

$$X(T) = \begin{cases} 0 \\ 1 \end{cases} 0 = \text{otherwise}$$

Savings possibilities include the following:

$$|x(t) = 1 | A(t) = Pr | E|x(t)| = X. Prx(t) = 1 \, t > 0 | E|x(t)| = X. Prx(t) = 1$$

As a static random, you'll also need to know the average on a real-time basis availability. The following is how the availability average is calculated:

$$A_{C = \frac{1}{C}} \int_0^C A(t) \, dt, \qquad c > 0$$

As a symbol denoting limited availability.

$$A = \lim_{t \to \infty} A(t)$$

The availability average on a real-time basis timeline is taken into account, along with field and randomization. As a result, the availability average can be represented as follows:

$$A_{\infty = \lim_{C \to \infty} A_C} = \lim_{C \to \infty} \frac{1}{C} \int_0^C A(t) \, dt \qquad c > 0$$

Protective monitoring of big data as periodic monitoring is critical to ensuring that cyberspace remains effectively protected, as we can eliminate vulnerabilities by identifying weak points and taking into account protection arrangements. The above equations have been applied to a sample of data. Since we can set availability at 96 percent and timeliness at 94 percent, systems can detect and prevent eavesdropping vulnerabilities, as well as control and respond appropriately.

Use a simple decision tree model for safety and security testing and verification: The basic decision tree model (String algorithm) is used to classify data, including input fields and variables shown in table no 1. The result of using this model to classify the higher education large data set. This template provides constrained reporting of database branches, such as classification or rejection decisions, as well as the scope for automatic error detection in big data.

Table 1. Description of Field name

Parameter	Description,
Input variable	x_1, x_2, x_3, x_4, x_5 & Y
Security rating	Security rating: 0= attack 1= security
Data risk	NO of test range of security 1=< 88.00 , 0 >88.00

The data input is coding correct(0, 1)&that dependant[x1-x2-x3-x4-x5-y].

The parse is located at the top of the first node of the tree shown in the image as infigure 5. It compiles data collection. We find in the typical dataset that is the safe protected data, with a rate of 54%, we can detect a rate of 4 percent and 12 degrees. It makes up 15% of the known risk and

vulnerability ratings for unprotected data. Since the initial stage in the study is to improve protection performance and security, let's see if each tree might provide any evidence of problematic components. The first split, as can be seen, depends on the level of the input data. As a result, it will be able to allocate or provide marks based on the maximum revenue of the class (2 knots). This rating has the highest percentage of unprotected data, indicating that data in this category is at risk and requires protection. As a result, the data in this category has a rate of 54 percent, indicating a risk, if not apparent.

Figure 5: Model of Decision Tree

Although the prediction model cannot respond in practice, we must ensure that the data is sufficiently fair enough to allow us to anticipate future risks and respond to potential threats for each degree based on the available data, just as if we analyzed the data referenced by Node 2 As can be seen, the vast majority (92308%) agree. It appears unprotected, which poses a risk and necessitates the implementation of a new security mechanism. As a result, we discovered that each score is an indicator of this model, and we may enhance security requirements in this data set to reduce risk. By allocating a specific node, we will be able to detect weak places.

Table 2. Descriptions of variables and parameters

	Through this stated, the following notation
N	The total number of instances observed
P	The number of parameters
Y	the jth case's the observed value of the dichotomous dependent variable
X	parameter n x 1 vector
B	Likelihood Function
W	log – likehold function
L	The total number of occurrences found.
L	The number of parameters is represented
I	the observed value

Preliminary Verifiability Control Calculations Using Logistic Regression Algorithms: The class A index is shown, followed by the Class B index X2 and the reference class IS C. Each score h passed through a logistic regression lateral model, the scores belonging to this category equally x1, x2, model set 0,0 which prospects for the threat that

each score h passed through a logistic regression lateral model, where each score h passes through A model prepared for logistic regression, the following formula is used to calculate the expected value as well as the degree of reliability: "most likely z = 1 value of scores and computation" is the expected value.

Table 2 shows that Possibility of ["flexible value feedback" &"Objectivity"&"accuracy verifiability"].

$$\hat{\pi}_i = Exp(\hat{\eta}_i)[\frac{1}{1 + Exp(\hat{\eta}_i)}]$$

$$wher \ \hat{\eta}_i \ as \ (X_i'\hat{\beta})$$

$IF \ \hat{\pi} > 0.5,$

In this case, the projected value is 1; otherwise, it is 0.

Confidence: The confidence value is (1-) for records having a projected value of y = 1, whereas it is (1+) for records with a predicted value of y = 0.

$(1-).\sum_{i=1}^{n} w_i \ (y_{i-} \pi_i)x_{ij} = 0$, for the parameter

Where $x_{i0=1} \ for \ i = 1, \dots, n.$

Table 3. Model fitting information

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	34.617			
Final	.000	34.617	8	.000

Pseudo R-Square

Cox and Snell	.750
Nagelkerke	1.000
McFadden	1.000

Is the usage of kind Newton Rafsson the Verifiability Algorithm? We can't rely on convergence, so we use Newton Rafsson-type algorithms to obtain the verifiability. We can't rely on closeness, so we use Newton Rafsson-type methods.• The estimated absolute difference between the frequencies.• The probability difference in percent.• Intervals between successive frequencies, the maximum number of frequencies specified to determine the validity of the data (verifiability) according to the frequency which is the lowest in all circumstances, and ranges from 10 to 8. The probability of redundancy occurring is very near to nil The message's frequency will be off, and all values will either be 1-0 or be released. Getting inverse of the first information matrix is the upper bound of the probability estimates and the variance matrix is the proxy estimated since the

previous. As well as Note: This example demonstrates how to use the total % as a design guide for the form. It might be a matter of precision. It might be the detective in certain circumstances. In general, the original; zero models were 72.6 percent accurate, Although the final model and expectation had a whole accuracy of 79.1%, as we saw, the real individuality of separate classes on a broad scale was among the forecasts' accuracy.

Table 4. expected values

Parameter Estimates

Wor(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
No	Intercept	-174.349	256116.563	.000	1	.999			
	Y	1.503	2034.026	.000	1	.999	4.494	.000	.(b)
	X1	.303	625.467	.000	1	1.000	1.354	.000	.(b)
	X2	-.406	1117.323	.000	1	1.000	.667	.000	.(b)
	X3	1.373	4523.229	.000	1	1.000	3.948	.000	.(b)
	X4	-1.175	2069.738	.000	1	1.000	.309	.000	.(b)
	X5	.026	1861.721	.000	1	1.000	1.026	.000	.(b)
	[Gra=1.000]	14.383	35793.597	.000	1	1.000	1764094.730	.000	.(b)
	[Gra=2.000]	33.400	54032.791	.000	1	1.000	320059604887103.000	.000	.(b)
	[Gra=3.000]	0(c)	.		0	.	.	.	.

a. The reference category is: Y.

b. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

c. This parameter is set to zero because it is redundant.

The output class with the highest value is the expected value for the scores. Opportunities with a logarithmic scale.

$$r_{ij} = \log\left(\frac{\pi_{ij}}{\pi_{ij}}\right) = x_i'\beta_j$$

$for\ j = 1, \ldots, j-1. The\ \log it\ for$ reference category $j, r, I, is\ 1.0.$

$$\hat{\pi}_{ij} = \frac{\exp(r_{ij'})}{1 + \sum_{k=1}^{i-1} \exp(r_{ij'})} = \frac{\exp(x_i'\beta_j)}{1 + \sum_{k=1}^{i-1} \exp(x_i'\beta_k)}$$

Table No. 3 shows that there are unlikely to be any changes for this classification when calculating the probability of occurrence of any of the previous criteria by calculating (exp), as we can see that "1-the lowest probability of occurrence is 0.390, which we find in the x4 variable, and the highest probability of occurrence is 4.494, which we find in the variable y 2- This model appears to be criteria. The outcomes of linking, feasible alternatives, and evaluating the likelihood of all j classes given comparable data are shown in Table 4.

CONCLUSION

Various educational bodies faced numerous challenges and obstacles, including network crimes, a lack of information technology governing security data, and the ability for education to create positive work climates by simplifying the processes for making necessary decisions with companies and reducing management steps to comply with regulatory obligations. Through the employment of a data mining model for the needed educational departments, artificial intelligence plays a significant role in monitoring the information analysis of e-learning systems.

It is recommended to analyze the data and look for hidden information in it. This technique, which uses association rules, basic decision trees, and Logistic Regression Algorithms as data mining tools, is a beneficial solution for a variety of threats that higher education institutions may face. maybe generalized; the suggested control strategy is quite thorough and lowers information security risks that impact senior management in educational institutions' strategic decisions. The findings of this study are highly valuable for developing a plan to assess the efficiency of learning performance to enhance the educational system's effective management, as well as the sort of data that will be employed, i.e. Finally, via real analysis discovered in dialogues, research, and blogging, this article may become an essential tool for learners in decision-making and monitoring of large data.

ACKNOWLEDGMENTS

REFERENCES

Aljahdali. T. H, Abdullah. S. A. L. (2020). Significance of Big Data Analytics in Higher Education Institutions of Saudi Arabia. Significance, 6(1).

Alkhalil. A Abdallah. M. A. E, Alogali. An Aljaloud. A, (2021). Applying Big Data Analytics in Higher Education: A Systematic Mapping Study. International Journal of Information and Communication Technology Education (IJICTE), pp. 29-51.

Al-Medlej. H. I, (1997). Decision-making process in higher education institutions: the case of Saudi Arabia, (Doctoral dissertation, Middlesex University).⁣

Alsheikh. N, (2019).Developing an Integrated Framework to Utilize Big Data for Higher Education Institutions in Saudi Arabia". International Journal of Computer Science & Information Technologypp. 31-33.

Aseeri. M, Kang. K, (2020). Technological and human factors for supporting big data analytics in Saudi Arabian higher education. In 26th Americas Conference on Information Systems, AMCIS 2020.

Baig. M. I, Shuib. L., Yadegaridehkordi. E, (2020). Big data in education: a state of the art, limitations, and future research directions. International Journal of Educational Technology in Higher Education, 17(1), pp. 1-23.

Bonilla. W, Al-Kamali. M, Farid. M, Mujahid. H, (2018).A business intelligence-based solution to support academic affairs: the case of Taibah University. Wireless Networks, 1-8⁣

Gandomi. A Haider. M, (2015). Big data concepts, methods, and analytics. International journal of information management, 35(2),pp. 137-144.

Hashim. F, Alam. G. M, Siraj. S, (2010). Information and communication technology for participatory-based

decision-making-E-management for administrative efficiency in Higher Education". International Journal of Physical Sciences, 5(4), pp. 383-392.

Janssen. M., van der Voort. H., Wahyudi. A, (2017). Factors influencing big data decision-making quality". Journal of business research, 70, pp. 338-345.

Kaiser. S, Armour. F, Espinosa J. A, and Money. W, (2013). Big Data: Issues and Challenges Moving Forward. 46th Hawaii International Conference on System Sciences, pp. 995-1004 DOI: 10.1109/HICSS.2013.645.

Meng. L. Q, Meng. L. Q. (2014). Application of Big Data in Higher Education. In 2nd International Conference on Teaching and Computational Science. Atlantis Press. (May 2014).

Mukthar. M. A, Sultan. M. M, (2017). Big Data analytics for higher education in Saudi Arabia. International Journal of Computer Science and Information Security (IJCSIS), 15(6).

Nda. R. M, Tasmin. R. B, (2019). Big Data Management in Education Sector: an Overview. Traektoriâ Nauki= Path of Science, 5(6).

Ong. V. K, (2015). Big data and its research implications for higher education: Cases from UK higher education institutions. In 2015 IIAI 4th International Congress on Advanced Applied Informatics, IEEEpp. 487-491.

Picciano. A. G, (2012). The evolution of big data and learning analytics in American higher education. Journal of asynchronous learning networks, 16(3), 9-20.

Segooa. M. A, Kalema. B. M, (2018). Leveraging big data analytics to improve decision-making in South African public universities. In 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), IEEE, pp. 9-13.

Segooa. M. A., Kalema. B. M. (2018). Improve Decision Making Towards Universities Performance Through Big Data Analytics. In the 2018 International Conference on Advances in Big Data, Computing, and Data Communication Systems (BCD), IEEE, pp. 1-5.

Verma. S, Bhattacharyya. S. S, Kumar. S. (2018). An extension of the technology acceptance model in the big data analytics system implementation environment. Information Processing & Management, 54(5), pp. 791-806.