JOURNAL OF AWARENESS
HOLISTENCE publications

Journal of Awareness
Volume / Cilt: 8, Issue / Sayı: 3, 2023, pp.297-303
E-ISSN: 2149-6544
https://journals.gen.tr/joa
https://doi.org/10.26809/joa.2081

Received / Geliş: 25.04.2023
Accepted / Kabul: 15.06.2023

RESEARCH ARTICLE/ARAŞTIRMA MAKALESİ

Application of clustering methods and data visualization for decision making in higher education

Olta LIaha[1] **Azir Aliu**[2] **Esmeralda Kadena**[3]

[1] PhD candidate., Faculty of Contemporary Sciences and Technologies, South East European University, Macedonia, e-mail: ol29064@seeu.edu.mk
[2] Prof., Faculty of Contemporary Sciences and Technologies, South East European University, Macedonia, e-mail: azir.aliu@seeu.edu.mk
[3] PhD., Bánki Donát Faculty of Mechanical and Safety Engineering, Obuda University, Hungary, e-mail: kadena.esmeralda@bgk.uni-obuda.hu

Abstract

This paper aims to present a case study to demonstrate the practical application of data visualization techniques and machine learning algorithms. Examining the data using various algorithms enables us to make predictions about how data visualization affects decision-making. Our study's results support the notion that data visualization influences decision-making. Moreover, we delve into the implications of employing data visualization technology in the academic community, including faculty and students. Furthermore, we evaluate how data visualization influences decision-making in higher education institutions, showcasing its potential to enhance the efficiency and speed of decision-making for stakeholders.

Keywords: Machine Learning, Higher Education, Clustering Methods, Data Visualization.

Citation/Atıf: LLAHA, O., ALIU, A. & KADENA, E. (2023). Application of clustering methods and data visualization for decision making higher education. *Journal of Awareness*. 8(3): 297-303, https://doi.org/10.26809/joa.2081

Corresponding Author/ Sorumlu Yazar:
Olta Llaha
E-mail: ol29064@seeu.edu.mk

1. INTRODUCTION

In today's information-driven world, the effective presentation of data in a communicative format has become crucial across all sectors. This growing trend of information visualization is driven by its ability to assist decision-makers in swiftly comprehending and analyzing vast amounts of information. By offering static and interactive visual representations of data, visualization serves as a powerful tool to enhance human understanding and facilitate data exploration and analysis. Users can interactively explore and analyze data, uncover significant patterns, infer correlations and causalities, and facilitate sense-making activities (M. O. Ward et al., 2015).

Institutions of higher education prioritize the creation of human capital and continuous improvement in quality through ongoing analysis. Key aspects for these institutions include predicting students' success and enhancing the work of academic staff. Machine Learning, a process in which computer systems automatically improve and learn from experience (Ayodele, Taiwo, 2010) plays a vital role in achieving these goals. Machine Learning empowers computers to discover insightful information without explicit guidance, utilizing algorithms trained on data to automate decision-making processes and iteratively generate outputs. Machine Learning research primarily focuses on generating models that illustrate the functioning of a system or concept, encompassing patterns, plans, representations, or descriptions, including mathematical operation rules and data patterns.

This paper is structured into the following sections:

Section 2 discusses the interplay between data visualization, machine learning, and higher education. Section 3 provides a comprehensive overview of the dataset and methodology employed. Section 4 presents the results obtained from the application of machine learning algorithms. Finally, section 5 concludes the study by discussing the implications of the findings.

2. MACHINE LEARNING AND DATA VISUALIZATION IN HIGHER EDUCATION

Assisting individuals in processing vast amounts of information while acknowledging the limitations of human cognitive capacities and the magnitude of data volumes is the primary objective of data visualization (Mohd, Maseri et al., 2010). To facilitate effective decision-making, it is crucial to design data visualization to allow domain experts to easily identify problems and solutions. Visualization techniques play a vital role in enhancing information comprehension and promoting visual thinking, making intricate relationships more accessible to understand (Shabdin, Nur et al., 2020).

In a case study conducted by (Klein, C et al., 2019) at a public university in the mid-Atlantic region of the United States, the use of learning analytics dashboard interventions by undergraduate students was examined. The study's findings revealed that the data's relevance, accuracy, and context significantly influence student understanding. Our own study further supports the notion that the utilization of data visualization greatly assists students in acquiring information. Specifically, our study focuses on data related to the application of data visualization technology, encompassing techniques, tools, and its impact on students' information acquisition. Furthermore, we explore its implications in classroom management, administration, and decision-making processes for academic staff.

Visualization techniques that leverage visual perception are particularly valuable in transforming data traces into meaningful visual information (Zotov, Vladimir et al., 2021). Data visualization enables comprehensive analysis of student assessments, facilitating informed decisions to enhance their performance. Moreover, it effectively communicates students' progress in various subjects and academic years by providing up-to-date assessments, comparisons with class averages, and predictions of future outcomes. Additionally, visually presenting educational issues based on students' responses yields more accurate information, potentially improving student behavior through

graphical observations.

3. METHODOLOGY

Data from a sample of participants, including academic staff and students, was collected through the administration of questionnaires, employing empirical methods. The diverse findings generated by the questionnaire responses are presented in Table 1, which depicts the sample selection process employed for the study.

Table 1. Sample selection for the study

Number of Respondents academic staff	Number of Respondents Students
38	78

The main objective of this research is to provide a comprehensive analysis of the current state of data visualization in higher education, with a specific focus on Logos University College. A questionnaire was utilized to assess the level of understanding and implementation of data visualization in decision-making processes. The questionnaire aims to measure the degree of familiarity and practical utilization of data visualization techniques.

A comparative analysis was conducted among various clustering methods to determine the most suitable clustering technique for this case study. The study involved the utilization of two datasets, and multiple clustering algorithms were implemented. These algorithms include SimpleKMeans, EM, MakeDensityBasedClusterer, FarthestFirst, Cobweb, HierarchicalClusterer, and FilteredClusterer. These algorithms were implemented using Weka 3.8.4, an open-source platform based on Java that integrates machine learning algorithms and data pre-processing tools. For analysis purposes, CSV data consisting of 38 observations and 16 variables for academic staff, as well as 78 observations and 13 variables for students, were selected. The attributes considered for the academic staff include Age, Gender, Academic qualification, Informed DV, Benefits DV, Data visualization affect teaching, DV affect classroom management and administration, DV help decision making, Data visualization technology, Data Visualization techniques, Data Visualization tools, and so on. Similarly, for the students, the attributes considered are Age, Gender, Program, Informed DV, Visualization affect information, Data visualization technology, Data Visualization techniques, Data Visualization tools, and so forth.

3.1. Clustering Algorithms

Clustering, a powerful data simplification and pattern discovery technique, involves partitioning data into groups of similar objects. While clustering may lead to a loss of certain details, it characterizes data based on its groupings and relies on mathematics, statistics, and numerical analysis. Clustering is a dynamic field of research within statistics, pattern recognition, and machine learning (Maimon Oded et al., 2010).

The K-means algorithm, described by (Jinxin Gao et al., 2009), is an iterative approach that aims to divide a dataset into K distinct and non-overlapping clusters. Each data point is exclusively assigned to one cluster. The algorithm maximizes similarity among data points within a cluster while ensuring maximum dissimilarity between clusters. It achieves this by minimizing the sum of squared distances between the data points and the cluster's centroid.

As explained by (Pankaj Saxena et al., 2017), the COBWEB algorithm generates a clustering dendrogram or a classification tree that represents each cluster probabilistically. This algorithm assumes attribute independence and employs hierarchical clustering to achieve predictability for nominal variable values.

The Farthest First algorithm, proposed by (Sapna Jain, et al., 2010), is a variant of the K-Means algorithm that initializes cluster centers using a different strategy. Instead of random selection, the Farthest First algorithm iteratively places each cluster center at the point farthest from existing cluster centers in the dataset. This approach aims to optimize clustering speed by minimizing reassignments and modifications during the process, improving efficiency in many cases.

EM (Expectation-Maximization), a widely used algorithm for partition methods (Mohammed J et al., 2014), reallocates instances from one group to another based on an initial partition. It seeks to minimize a specific error criterion and is specifically designed for compact and distinct groups. EM clustering generates probabilistic descriptions of clusters, including mean and standard deviation for numerical attributes and adjusted value counts (to prevent zero probabilities) for nominal attributes.

The Make Density-based Cluster algorithm, introduced by (Sapna Jain, et al., 2010), is a significant clustering algorithm capable of effectively identifying clusters of arbitrary shapes and handling noise. One of its advantages is that it requires only a single pass through the raw data. This algorithm defines clusters as dense regions separated by areas with lower object density, making it suitable for clustering data streams. Density-based clustering identifies clusters based on the minimum number of neighboring instances within a specified radius, helping to determine clusters based on data point density in the vicinity.

4. EXPERIMENTAL RESULTS

Presented in Table 2 and Table 3 are the outcomes of the analysis conducted using seven clustering algorithms: SimpleKMeans, MakeDensityBasedClusterer, EM, Cobweb, HierarchicalClusterer, FarthestFirst, and FilteredClusterer. These algorithms were assessed based on the similarities among objects and the time required for cluster generation. In this case study, the evaluation option chosen for both datasets is the classes to cluster evaluation, specifically focusing on the attribute of Visualization affecting/helping decision making.

Table 2 and Table 3 showcase the diverse findings obtained from applying these clustering techniques. The intention behind sharing these results is to provide a model that can support higher education institutions in improving their decision-making processes and optimizing strategic objectives.

Table 2. Results of Clustering Techniques for Academic Staff

Algorithm	No. of Clusters	No. of Iteration	Log Likelihood	Sum of squared errors
EM	3	0	-12.79998	
Cobweb	39			
HierarchicalClusterer	1			
SimpleKMeans	2	3		184.0
MakeDensity BasedClusterer	2	3	-13.84669	184.0
FarthestFirst	2			
FilteredClusterer	2	3		184.0

Comparable clustered instances are generated by the SimpleKMeans, MakeDensityBasedClusterer, FarthestFirst, and FilteredClusterer clustering algorithms, while the ey, Cobweb, and HierarchicalClusterer algorithms produce distinct results. It is worth mentioning that the MakeDensityBasedClusterer and EM algorithms exhibit higher likelihood values, indicating the presence of more reliable clusters.

Table 3. Results of Clustering Techniques for Students

Algorithm	No. of Clusters	No. of Iteration	Log Likelihood	Sum of squared errors
EM	3	10	-10.7833	
Cobweb	89			
HierarchicalClusterer	1			
SimpleKMeans	2	6		270.0
MakeDensity BasedClusterer	2	6	-11.2990	270.0
FarthestFirst	2			
FilteredClusterer	2	6		270.0

In contrast to the EM, Cobweb, and HierarchicalClusterer algorithms, the SimpleKMeans, MakeDensityBasedClusterer, FarthestFirst, and FilteredClusterer algorithms yield comparable clustered instances. MakeDensityBasedClusterer and EM demonstrate higher likelihood values within this set of algorithms, suggesting a greater probability of generating dependable clusters.

Table 4. Performance Comparison – Clustering for Academic Staff

Algorithm	Incorrectly clustered instances (%)	Time taken to build (seconds)
EM	36.84	0.37
Cobweb	68.42	0.06
HierarchicalClusterer	5.26	0.04
SimpleKMeans	31.58	0
MakeDensity BasedClusterer	31.58	0.01
FarthestFirst	28.95	0.02
FilteredClusterer	31.58	0

Table 4 presents information regarding the duration required to construct a cluster and the count of incorrectly clustered instances. The algorithms EM, Cobweb, and HierarchicalClusterer demonstrate longer construction times and exhibit a higher number of incorrectly clustered instances. Conversely, the algorithms SimpleKMeans, FarthestFirst, and FilteredClusterer outperform the other algorithms in terms of cluster quality. A greater number of incorrectly clustered instances indicates inferior cluster quality. Consequently, the SimpleKMeans and FilteredClusterer algorithms are considered superior due to their shorter construction times and a reduced number of incorrectly clustered instances.

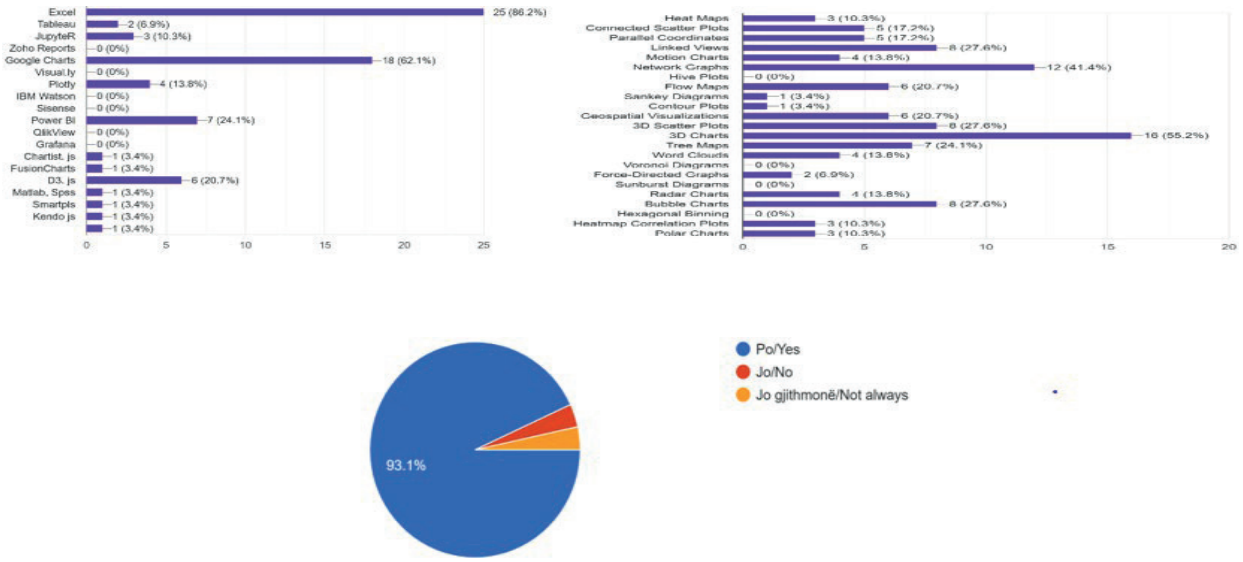
In relation to the students, the class attribute encompasses Data visualization, which impacts the information.

Table 5. Performance Comparison – Clustering for Students

Algorithm	Incorrectly clustered instances (%)	Time taken to build (seconds)
EM	46.75	0.42
Cobweb	70.13	0.06
HierarchicalClusterer	41.56	0.06
SimpleKMeans	42.86	0.01
MakeDensity BasedClusterer	44.16	0.01
FarthestFirst	38.96	0.01
FilteredClusterer	42.86	0

Table 5 presents a summary of the time required to construct a model and the count of instances that were incorrectly clustered. EM, Cobweb, and HierarchicalClusterer exhibit longer construction times among the algorithms. Furthermore, incorrectly clustered instances are higher with the EM, Cobweb, and MakeDensityBasedClusterer algorithms. Consequently, the SimpleKMeans, MakeDensityBasedClusterer, HierarchicalClusterer, and FilteredClusterer algorithms showcase superior performance compared to the rest. It is worth emphasizing that many incorrectly clustered instances do not signify a reliable cluster. Hence, the FilteredClusterer algorithm emerges as the optimal choice due to its shorter construction time and reduced count of incorrectly clustered instances.

Figure 1. Data visualization for academic staff



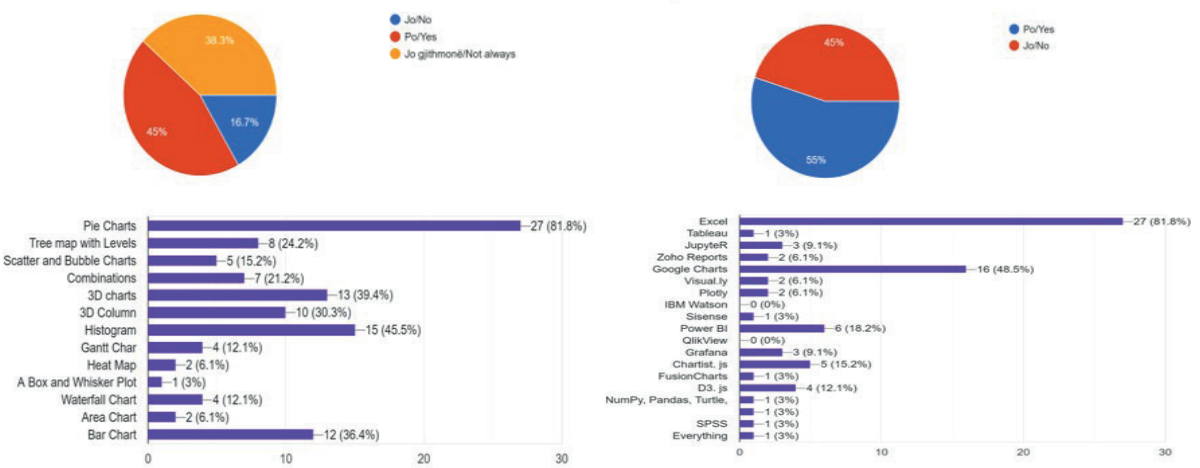
When it comes to incorporating visualization into teaching, the staff encounters specific challenges. These challenges encompass identifying suitable visualizations, selecting appropriate data, ensuring the data is presented clearly, and maintaining students' engagement throughout the process. Furthermore, 81.1% of the staff reports that visualization significantly impacts classroom management and administration. Excel stands out as the staff's most frequently utilized tool for data visualization, with a usage rate of 86.2%. Tableau and Google Chart follow closely behind with a usage rate of 62.1%, while Power BI is used by 24.1% of the staff. It is worth mentioning that 93% of the staff considers data visualization to be crucial for decision-making. Valuable techniques for data visualization encompass 3D charts, tree maps, bubble charts, geospatial visualizations, and various others. Figure 1 visually represents these notable findings.

The positive impact of data visualization on students' understanding of information is clearly evident, as attested by 83.3% of the surveyed students. Out of these students, 55% actively make use of data visualization technology. They employ a diverse range of visualization techniques, including Pie charts, Histograms, Bar charts, and 3D charts, among others. In conjunction with these techniques, students utilize a variety of tools for data visualization, such as Excel, Google charts, D3.js, and Power BI, as illustrated in Figure 2.

5. CONCLUSIONS

The study extensively delves into the role and significance of data visualization, emphasizing its crucial importance for individuals handling large and intricate datasets. The survey results confirm the considerable impact of data visualization on data analysis, comprehension, and decision-making. Moreover, the study reveals the continuous evolution of techniques and methods for visual data representation, parallel to advancements in technology and theoretical knowledge. As a valuable contribution to the field of data visualization in higher education, this study involves designing and creating a dataset specifically utilized for data visualization purposes employing diverse techniques. The primary objective of data visualization is to aid individuals in interpreting data and making well-informed decisions based on it. Within higher education, data visualization can enhance data utilization by various stakeholders, including educators and administrators, in decision-making processes related to student performance, course registration, and schedule creation. In this particular case study, the main aim was to identify the most effective clustering algorithms for predicting the impact of data visualization technology on decision-making. To accomplish this, a comparison was made among different clustering algorithms based on criteria such as execution time, number of iterations, and instances incorrectly clustered. Additionally, machine learning algorithms were implemented

Figure 2. Data visualization for students



and their performance was evaluated to predict the usage of data visualization and its influence on decision-making. The application of data visualization in higher education institutions exhibits a demonstrable impact on decision-making by enabling stakeholders to make faster and more well-informed decisions, an impact that was thoroughly evaluated in this study.

REFERENCES

AYODELE, T. (2010). *Machine Learning Overview*. 10.5772/9374

JİNXİN G., DAVİD B. HİTCHCOCK JAMES-STEİN. (2009) Shrinkage to Improve K-means Cluster Analysis University of South Carolina, Department of Statistics November 30, 2009

KLEIN, C., LESTER, J., NGUYEN, T., JUSTEN, A., RANGWALA, H., & JOHRI, A. (2019). Student Sensemaking of Learning Analytics Dashboard Interventions in Higher Education. *Journal of Educational Technology Systems*, 48(1), 130–154. https://doi.org/10.1177/0047239519859854

MAIMON, O. & ROKACH, L. (2010) Data mining and knowledge discovery handbook Chapter 15 clustering method. http://www.ise.bgu.ac.il/faculty/liorr/hbchap15.pdf

MOHD, M., EMBONG, A. & MOHAMAD ZAIN, JASNI. (2010). A Framework of Dashboard System for Higher Education Using Graph-Based Visualization Technique. 87. 55- 69. 10.1007/978-3-642-14292-5_7

MOHAMMED J. Z., & WAGNER MEIRA, JR. (2014) *Data Mining and Analysis Fondamental Concepts and Algorithms, ISBN: 978-0-521-76633-3*

PANKAJ, S, & SUSHMA L. *(2017) Analysis of Various Clustering Algorithms of Data Mining on Health Informatics, International Journal of Computer & Communication Technology ISSN (PRINT): 0975 -7449, Volume 6, Issue-2, 2017*

SAPNA J., M AFSHAR A. & M N DOJA. (2010) K-means clustering using weka interface, *Proceedings of the 4th National Conference*; INDIACom-2010

SHABDIN, N., YAACOB, SURAYA & SJARIF, N.N.A. (2020). *Relationship Types in Visual Analytics*. 1-6. 10.1145/3397125.3397127

WARD, M. O., GRINSTEIN, G. & D. KEIM. (2015) *Interactive Data Visualization: Foundations, Techniques, and Applications*, Second Edition. A. K. Peters, Ltd., 2015.

ZOTOV, V., IBRAHIM, I. & PETUNINA, I. & LAZAREVA, Y.. (2021). Engagement of Students in Data Visualization for the Purpose of E-Learning Improvement. International *Journal of Emerging Technologies in Learning (iJET)*. 16. 46. 10.3991/ijet.v16i02.18745