INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 4, PP. 405-412
Manuscript received December 26, 2023; revised June, 2024. doi: 10.24425/ijet.2024.149559

Electronic footprint analysis and cluster analysis techniques for information security risk research of university digital systems

Valerii Lakhno, Myroslav Lakhno, Kaiyrbek Makulov, Olena Kryvoruchko, Alona Desiatko,Vitalii Chubaievskyi, Dmytro Ishchuk and Viktoriya Kabylbekova

***Abstract*—In the article there are presented results of the study of the state of user competencies for different specialties of the university digital educational environment (UDEE) on issues related to information security (IS). The methods of cluster analysis and analysis of digital (electronic) traces (DT) of users are used. On the basis of analyzing the DTs of different groups of registered users in the UDEE, 6 types of users are identified. These types of users were a result of applying hierarchical classification and k-means method. Users were divided into appropriate clusters according to the criteria affecting IS risks. For each cluster, the UDEE IS expert can determine the probability of occurrence of high IS risk incidents and, accordingly, measures can be taken to address the causes of such incidents. The algorithms proposed in this study enable research during log file analysis aimed at identifying breaches of information security within the university's DEE.**

***Keywords*—cluster analysis; digital footprints; hierarchical classification; k-means method; information security; risks; university digital educational environment**

I. Introduction

RECENTLY, most educational institutions around the world, especially against the backdrop of the rapid spread of the Covid-19 pandemic, offer various forms of online learning to students in addition to traditional forms of learning. Such a format of learning, for example, is realized through distance learning systems (DLS) or online courses in specific disciplines. However, for the effective functioning of such online courses and DLSs in general, an appropriate digital environment or information system (IS) is needed to perform the functions of managing and organizing the learning process. In most cases, such tasks are solved with the help of well-proven learning management systems - LMSs, which offer a variety of functions to support teachers in the process of creating, administering, and managing online courses.

However, in most cases such systems have a rather limited arsenal of modules for analyzing data collected during the learning process [1, 2]. Such systems do not sufficiently realize the potential for analyzing the so-called digital traces (DT) of students during the study period [3]. In addition, such LMS systems practically do not use the accumulated data on students' digital traces to increase the degree of protection and information security of the university from external intrusions.

In the university digital educational environment (hereinafter referred to as UDEE), students' DTs can be related to the technologies of intelligent data analysis (IDA) in various aspects. According to several authors, DT and IDA technologies can be used for [4, 5]:

- improving the quality of educational processes at the university. This is because analyzing DT with the help of IDA technologies helps to establish dependencies between such parameters as: attendance data; activity on online platforms; current and final course grades; etc. Accordingly, DT and IDA technologies can be used to identify topics and modules in which students are struggling and can assist educators in choosing solutions to provide personalized support and guidance to specific students;
- predicting academic performance. Analysis of students' DT can help in predicting their academic performance. Thus, with the help of IDA technologies it is possible to study the relationships between various factors, including students' activity in classes, participation in discussions, deadlines for assignments, course grades, etc. This analysis will help to determine which factors can affect the success of students and predict their future academic performance. Such analysis will eventually help to determine which factors can more or less influence the success of students' learning and predict their future performance in other disciplines during, for example, a semester;
- decision support. The analysis of students' DT can be used to support decision-making by the university administration. For example, based on the analysis of data obtained during DT analysis, university

Valerii Lakhno and Myroslav Lakhno are with National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine (e-mail: lva964@nubip.edu.ua, lvaua21@gmail.com)

Kaiyrbek Makulov and Viktoriya Kabylbekova are with Caspian University of Technology and Engineering named after Sh.Yesenova, Almaty, Kazakhstan (e-mail: kaiyrbek.makulov@yu.edu.kz, viktoriya.kabylbekova@yu.edu.kz)

Olena Kryvoruchko, Alona Desiatko and Vitaliy Chubaievskyi are with State University of Trade and Economics, Kyiv, Ukraine (e-mail: ev_kryvoruchko@ukr.net, desyatko@gmail.com, chubaievsлyi_vi@knute.edu.ua)

Dmytro Ishchuk is with Zhytomyr Politechnic State University, Zhytomyr, Ukraine (e-mail: o.ishchuk@knute.edu.ua)

management can determine the effectiveness of educational programs; identify potential problems, as well as suggest ways to improve the educational process. In addition, such analysis of DT with the help of IDA methods contributes to the selection of effective methods of teaching and assessment of students' work, which will eventually allow to determine the most effective strategies and approaches for the implementation of the educational process;

Personalized learning: Analysis of students' DT can help to create personalized learning programs. Based on data about students' past successes and preferences, individualized learning plans can be developed that are tailored to their needs and interests. This will allow students to receive a better education.

Thus, IDA technologies play an important role in processing and analyzing large amounts of data related to various student activities, allowing universities to obtain valuable information to improve the quality of the educational process and support students in the process of forming their individual educational trajectory.

Note that also in the context of information security (IS) of the UDEE, IDA and DT technologies and methods can also be useful, e.g., for:

- monitoring student and faculty network activity. IDA techniques can help identify anomalies and unusual activity that typically indicate unauthorized access (UA) attempts or hacking of the university network;
- analysis of student and faculty authorization and authentication logs. Accordingly, the use of IDA techniques and DT analysis can detect unusual or suspicious attempts to log in to or authorize at the UDEE. Thus, such analysis can assist in the identification of UA attempts or hacking of UDEE user accounts;
- detection of malware and attacks on UDEE;
- analyzing the IS threats to the UDEE in general. IDA and DT methods can be used to analyze and classify IS threats, such as phishing attacks, buffer overflow attacks, denial of service attacks, etc.;
- vulnerability prediction. The use of prediction methods, IDA and DT analysis allows predicting possible vulnerabilities and weaknesses in the UDEE IS loops. Thus, it helps in taking measures to prevent vulnerabilities and improve the IS level of the UDEE.

It is important to note that the successful implementation of methods, algorithms and appropriate software for IDA and DT analysis in the context of ensuring the security of the UDEE, as well as improving the quality of the educational process, requires the collection and processing of large amounts of data, as well as the use of appropriate algorithms and machine learning models. In addition, legal and ethical considerations must be taken into account when using such methods to ensure the privacy and data protection of UDEE users.

All the above mentioned predetermines the relevance of this study and motivates us to continue the work in this direction.

II. LITERATURE REVIEW

In [4], the authors show that the study of learning processes can benefit from learning the DTs that students leave in particular when browsing educational platforms. In their paper, the authors describe a process detection experiment based on students moving paragraphs of web pages of a Python tutorial. Such observation of DT was intended to study students' interaction with the discipline and related disciplines in terms of the concept of instructional coherence. However, the work is experimental in nature without detailing the choice and validity of the methods used in the process of analyzing DT.

In [5], the authors present the results of research on the integration of learning styles based on DT analysis into web-based educational systems on LMS Claroline, Ganesha, Chamilo, and Moodle. This integration of DT analysis results, according to the authors, will help students in course mastery. However, the authors limited themselves only to a comparative analysis of LMS systems, without detailing the methods of applying a particular method to analyze students' DTs in LMS system.

In [6], the authors, resorting to the methods of cluster analysis of data, obtained also from students' digital footprints, tried to solve the problem related to the ways to improve the efficiency of the educational process. However, the work does not address many technological aspects of analyzing students' DTs.

In [7], the authors of the study, resorting to the use of correlation and cluster analysis methods, analyze the strategies of educational activities of students in various social networks.

In [8], [9], [10], the authors investigated in detail using various IDA methods the dependencies between students' successes and their LMS Moodle activities. This kind of research allowed to identify the behavioral strategies of students during online learning.

A separate and rather large array of research publications is devoted to the application of data mining methods to increase the degree of security of various objects of information activity.

The dynamics of growth of IS incidents in educational institutions that have emphasized the introduction of information technologies and systems into the educational process shows the evolution of this type of threats.

In [11], [12], [13], [14], [15], the authors address various aspects of the problems of applying IDA techniques for information security (IS) of UDEE.

Thus, in [11] the study relied on the analytical-descriptive approach in the analysis of intellectual security in the work of organizational committees working on educational programs. However, the authors of this study focused more on intellectual security itself. At the same time, minimal attention is paid to the issues of providing IS data in educational programs.

It is shown in [12], [13] that the use of IDA techniques can be very effective in addressing the problems related to intrusion detection and privacy protection.

The work [14] is devoted to the identification of IS threats to a university campus based on the use of IDA methods. However, the authors limited themselves to only a small list of possible threats, without addressing the issue of the relationship between threats and students' DTs.

In [15], an approach of applying IDA techniques for IS risk assessment is discussed.

In works [16], [17], [18], the authors touch upon the problem of research of the bundle - "Digital footprint"-"Intelligent data analysis"-"Information security".

Thus, [16] shows that IDA technologies can be successfully used to track the spread of malware, identify the sources of

cyberattacks, and understand the motives and tactics of cybercriminals.

Thus in [17], the authors review studies on various factors that influence people's willingness to leave DTs on social networks, which can be useful for investigating factors related to users' IS, such as UDEE.

Thus, [18] explores the implications of the trade-off between digital footprint privacy protection and commercial applications, using marketing as an example. The authors propose a theoretical and practical mechanism for protecting the privacy of digital footprints.

Among the publications on the application of IDA methods to analyze DT, we can single out the works devoted to cluster analysis (CA). From our point of view, this is because the CA method can be useful for analyzing the DTs of students in UDEE. For example, CA methods allow to perform grouping of students or their actions, in the context of ensuring the IS of UDEE, based on their behavior and other parameters. This allows you to identify similar patterns of student behavior, identify problem areas. In addition, cluster analysis can help to personalize learning, allowing more effective tailoring of training programs to the needs of each group of students.

Thus, [19] addresses the problem of identifying inaccuracies or non-obvious phenomena among cybercrime using CA methods. The work explores selected technical, contextual, and behavioral characteristics that can be routinely selected in the process of clustering cybercrime features.

In [20], digital tracing programs are analyzed, which is one of the restrictive measures in the implementation of the COVID-19 pandemic containment strategy. According to the authors, DT analysis is an alternative to traditional contact tracing. The focus of the paper is on the use of cluster analysis of some factors concerning the possible behavior of respondents.

As the analysis of previous studies has shown, cluster analysis (CA) can be a useful tool for dividing students enrolled in UDEE and in DLS by the level of their technical knowledge about IS, risks and measures for safe work in UDEE. Cluster analysis methods will allow to identify groups of students based on similar characteristics and needs, which can also be used in the formation of individual educational programs.

The attributes that can be identified from analyzing students' DTs to form clusters may include:

- - the level of IS knowledge. Accordingly, different aspects of IS are considered, such as basic IS concepts, awareness of the current IS threat landscape and protection methods;
- - experience with computer systems and networks. Skills in operating systems (OS), system and application software, and knowledge of networking technologies are taken into account;
- - knowledge of basic principles of secure networking. These attributes include: knowledge of password rules, use of anti-virus software, awareness of social engineering techniques used by hackers to break in, knowledge of phishing techniques.

From the point of view of ensuring the UDEE IS, a relative difficulty in using student clustering methods is to identify groups of students (and possibly teachers) whose working style in the UDEE may pose a threat to the UDEE IS or carry direct threats to the IS of the university in general. That predetermined the purpose of this study.

In summary, it can be stated that the development of models and methods of intelligent data analysis concerning the digital traces of students and educators in the digital educational environment (DEE) of educational institutions is a relevant task, and some aspects of its solution are the subject of this work.

III. The purpose of the study.

The purpose of the study is to identify the groups in the UDEE with the lowest context-dependent characteristics related to IS compliance issues when working in the UDEE and the levels of IS risks when different categories of users work in the UDEE.

In this regard, such challenges need to be resolved:

1) identify an array of similar groups according to certain criteria;

2) analyze the DT of the students assigned to this group and conduct cluster analysis based on the k-means method.

IV. Methods and models

A CA and DT analysis is used to identify groups of students and faculty at UDEE with the least knowledge and experience in IS issues. The following steps were sequentially implemented:

Step 1 - Data Collection. Available data of users registered in the UDEE were collected using the relevant LMS Moodle and OS logs. These data described, for example, the completion of relevant IS courses, the extent and success rate (for students) of IS assignments, test results, etc. for IS related courses. Also, these data were obtained from: the assessment of students' test results on IS topics; monitoring of users' online activity; assessment of security measures at the UDEE (e.g., how actively students use security measures such as strong passwords, two-factor authentication, and data encryption, which indicates users' awareness of IS threats);

Step 2 - Data Preparation. The collected data were filtered to remove incomplete records. And then converted into numerical format.

Step 3 - Analyze digital footprints: DTs of UDEE users may indicate IS risks. Such DTs can include: Behavioral data (Unusual attempts to access, for example, the DLS, frequently failed account login attempts, unexpected changes of location or devices to access learning materials; Abnormal activity (Unusually heavy downloading or copying of materials beyond normal use); Deviations from normal schedules (Unexpected or unusual periods of activity, logging in at unusual times for a particular student); Changes in normal patterns of behavior; Unusual requests to change credentials. Using DT analysis techniques, additional information about students and faculty members' IS cognition and experience was also extracted. For example, this could be an analysis of activity in UDEE, behavioral style, participation in discussion forums, etc.

Phase 4 - Implementation of CA. Applying CA methods to group students based on their knowledge and experience in IS.

Step 5- Interpretation of results.

Note that the results of the CA and DT analysis may be approximate and may require additional interpretation and verification. However, these methods can help the institution's management to begin the process of identifying groups of

students with the least knowledge and experience in IS issues to improve the security of the UDEE as a whole.

Let's denote by Ω the whole set of users registered in the UDEE. To classify and cluster Ω we need to form homogeneous clusters of users according to the degree of risk to the UDEE IS. For this purpose, we will use the methods of cluster analysis - hierarchical classification and k-means method.

In the simplest variant the Euclidean distance can be used as a metric

$$\rho_{ik} = \sqrt{\sum_{i=1}^{m}(z_{ji} - z_{jk})^2}, \quad (1)$$

Where ρ_{ik} - is the distance between, respectively, i, k observations generated from the standardized data mentioned above;

z_{ji} – standardized data matrix.

However, this normalization of data will cause items with large absolute values and standard deviations to have a large impact on items with smaller absolute values and standard deviations [21].

Let's perform clustering using the closest possible approximation of the user groups registered in the UDEE.

Let the set Ω contains such six groups, see Table 1. The data were obtained based on the analysis of DT of users of different specialties for students of two universities - National University of Bioresources and Nature Management of Ukraine and Esenov University (Kazakhstan, Aktau).

TABLE I
CLASSIFICATION OF USERS REGISTERED IN THE UDEE BY IS RISK CRITERION

№	Group name (types of users)	Traits (Description of Behavioral Pattern)
1	Informed users	This group includes users who are aware of the risks on the university network (or the UDEE as a whole). These users take measures to ensure the security of their data and accounts in the UDEE. They always follow the recommendations for creating complex passwords, regularly update software, do not open suspicious links or attachments in emails, and use reliable anti-virus software on their PCs.
2	Careless users	The group includes users who do not pay due attention to IS measures. Such users are vulnerable to attacks. Users in this group are characterized by using weak passwords, repeating passwords for different accounts, not updating software in a timely manner, ignoring suspicious activity and measures to protect their data in the UDEE.
3	Ignorant users	The group contains users who have insufficient knowledge of IS risks when working on the university network. This group may use insecure software on their PCs when connecting to the UDEE, and often transmit sensitive data through insecure communication channels.
4	Indifferent users	The group includes users who are not interested in the UDEE IS issues. This category of users does not check their accounts for hacking, does not pay attention to warnings about possible threats, etc.
5	Irresponsible users	Participants violate IS rules and policies on the University network. They may attempt to gain unauthorized access to UDEE systems, distribute malware, violate data privacy, or engage in unscrupulous activity at UDEE.
6	Destructive users	These users, attempt to cause damage to the University network, including spreading viruses, blocking network resources, etc.
Note. These user groups are rather conventional. There are often no clear boundaries between user groups. There may also be nuances between groups. As users gain knowledge, for example through appropriate courses, they may move from one type to another, recognizing the importance of IS on the network and taking appropriate measures to protect their data and accounts.		

Any user group can be characterized using the following attributes, see Figures 1, 2 (Figure 2 is a table with standardized variables):

Name of the specialty and user group ID (NS_IDGroup - taken, for example, from LMS Claroline, Ganesha, Chamilo, Moodle);

Year of study 1-5 (Year_ study - e.g. 1 (1st year) etc.) (taken e.g. from LMS Claroline, Ganesha, Chamilo, Moodle));

Personal responsibility (Pers_ resp (1-100) - characterized by the presence of specific attributes, e.g., the degree of accuracy in processing personal data, frequency of changing passwords, basic knowledge of IS, activity of using secure connections when working in the DLS, etc. - these data can be partially obtained on the basis of user DT analysis using LMS Claroline, Ganesha, Chamilo, Moodle, as well as using SIEM, e.g., Splunk, see (3). 3);

Average performance in IT and IS related courses (Aver_ perf (0-100) - taken e.g. from LMS Claroline, Ganesha, Chamilo, Moodle));

IS competency assessment (Comp_ ass (0-10), e.g. knowledge of: phishing; malware; use of weak/strong passwords; software updates, including operating system; network security, including use of secure Wi-Fi; basics of firewall configuration and VPN use; secure data storage, etc., e.g. taken from testing and/or questionnaires of UDEE users).

.

NS_IDGroup	Year_ study	Pers_ resp	Aver_ perf	Comp_ ass
Software Engineering_1501	1	21	68	4
Software Engineering_1501	2	40	74	5
Software Engineering_1501	3	56	77	6
Software Engineering_1501	4	65	84	6
Information Technology_1502	1	17	65	4
Information Technology_1502	2	37	77	5
Information Technology_1502	3	61	81	6
Information Technology_1502	4	71	96	6
Economic cybernetics_1503	1	21	71	4
Economic cybernetics_1503	2	44	76	5
Economic cybernetics_1503	3	60	80	6
Economic cybernetics_1503	4	71	74	7
Computer engineering_1504	1	22	69	4
Computer engineering_1504	2	42	75	5
Computer engineering_1504	3	61	74	7
Computer engineering_1504	4	75	78	9
Economy_1601	1	14	61	2
Economy_1601	2	29	73	3
Economy_1601	3	45	74	4
Economy_1601	4	50	75	4
Enterprise economy_1602	1	12	62	2
Enterprise economy_1602	2	25	73	3
Enterprise economy_1602	3	40	73	4
Enterprise economy_1602	4	49	75	4
Management_1701	1	11	65	2
Management_1701	2	27	72	3
Management_1701	3	39	73	3
Management_1701	4	42	75	4
Ecology_1604	1	12	66	2
Ecology_1604	2	26	73	3
Ecology_1604	3	34	72	3
Ecology_1604	4	41	80	4

Fig. 1. Screenshot of the table in STATISTICA 12.5 with raw data for analysis

NS_IDGroup	Year_ study	Pers_ resp	Aver_ perf	Comp_ ass
Software Engineering_1501	-1,32051	-0,97789	-0,87045	-0,20733
Software Engineering_1501	-0,44017	0,033262	0,032936	0,395806
Software Engineering_1501	0,44017	0,884758	0,484629	0,998939
Software Engineering_1501	1,320511	1,363725	1,538579	0,998939
Information Technology_1502	-1,32051	-1,19077	-1,32214	-0,20733
Information Technology_1502	-0,44017	-0,12639	0,484629	0,395806
Information Technology_1502	0,44017	1,150851	1,086886	0,998939
Information Technology_1502	1,320511	1,683037	3,345352	0,998939
Economic cybernetics_1503	-1,32051	-0,97789	-0,41876	-0,20733
Economic cybernetics_1503	-0,44017	0,246136	0,334065	0,395806
Economic cybernetics_1503	0,44017	1,097633	0,936322	0,998939
Economic cybernetics_1503	1,320511	1,683037	0,032936	1,602073
Computer engineering_1504	-1,32051	-0,92467	-0,71989	-0,20733
Computer engineering_1504	-0,44017	0,139699	0,1835	0,395806
Computer engineering_1504	0,44017	1,150851	0,032936	1,602073
Computer engineering_1504	1,320511	1,895911	0,635193	2,808339
Economy_1601	-1,32051	-1,35042	-1,9244	-1,41359
Economy_1601	-0,44017	-0,55214	-0,11763	-0,81046
Economy_1601	0,44017	0,299354	0,032936	-0,20733
Economy_1601	1,320511	0,565447	0,1835	-0,20733
Enterprise economy_1602	-1,32051	-1,45686	-1,77384	-1,41359
Enterprise economy_1602	-0,44017	-0,76502	-0,11763	-0,81046
Enterprise economy_1602	0,44017	0,033262	-0,11763	-0,20733
Enterprise economy_1602	1,320511	0,512229	0,1835	-0,20733
Management_1701	-1,32051	-1,51008	-1,32214	-1,41359
Management_1701	-0,44017	-0,65858	-0,26819	-0,81046
Management_1701	0,44017	-0,01996	-0,11763	-0,81046
Management_1701	1,320511	0,139699	0,1835	-0,20733
Ecology_1604	-1,32051	-1,45686	-1,17158	-1,41359
Ecology_1604	-0,44017	-0,7118	-0,11763	-0,81046
Ecology_1604	0,44017	-0,28605	-0,26819	-0,81046
Ecology_1604	1,320511	0,08648	0,936322	-0,20733

Fig. 2. Screenshot of the table with standardized variables

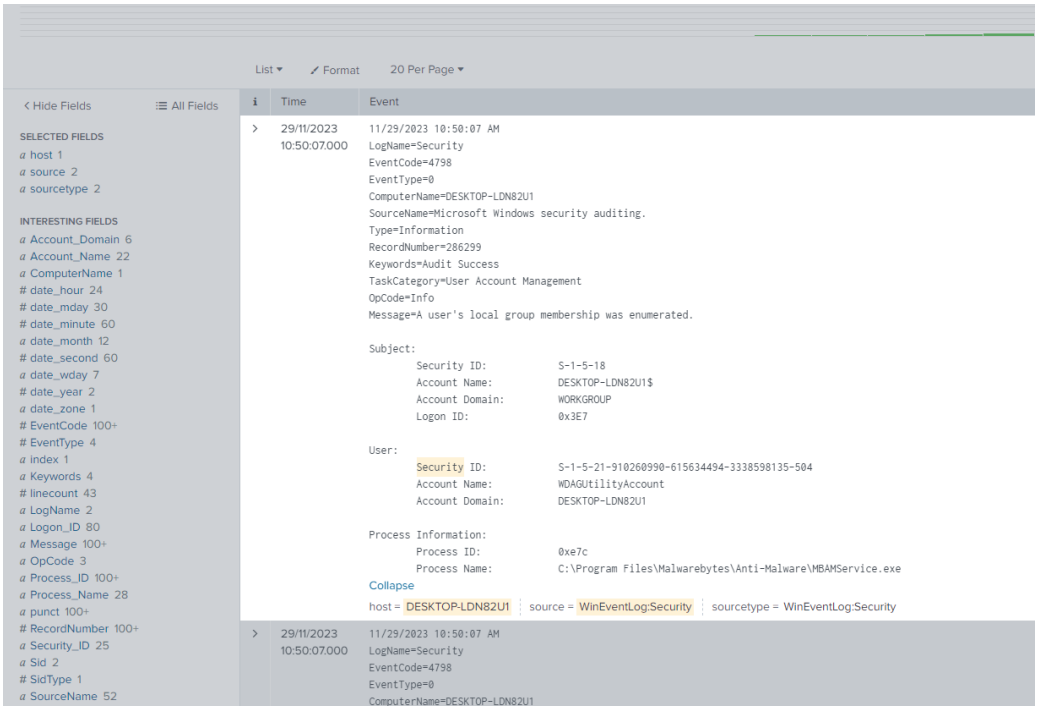

Fig. 3. Using SIEM Splunk to collect and analyze digital footprints

V. AN EXPERIMENTAL STUDY

The purpose of the CA is to divide the users of the UDEE into classes. Each of these classes corresponds to its own risk group in the context of IS. Observations falling into the same group are characterized by the same probability of an IS incident.

The study was conducted using the STATISTICA 12.5 package

At the first stage, the dendrograms necessary for grouping objects into subsets (clusters) based on their similarities were obtained based on hierarchical classification, see Fig. 4.

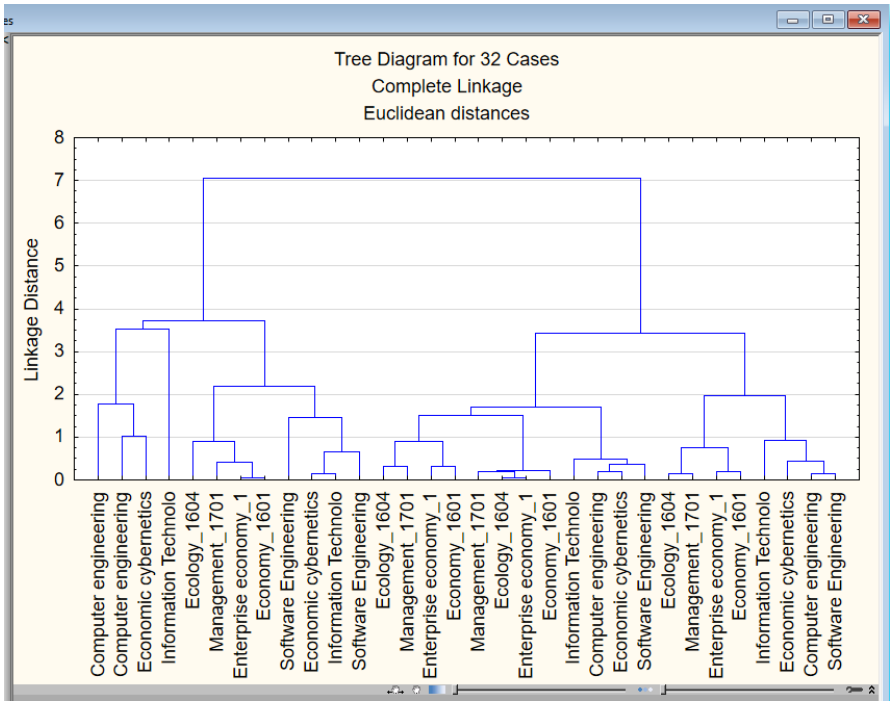

Fig. 4. Screenshot of hierarchical classification results (STATISTICA 12.5 package)

It is necessary to identify appropriate clusters among the existing set of groups. For example, users classified as dangerous in terms of IS criteria and risks to the UDEE, indifferent users and neutral users according to the assessed attributes.

Assumption. Data on all the above attributes is available and can be assessed and measured. For example, password changes, IP changes and other digital footprints can be tracked in LMS systems (see Figure 5) as well as with the previously mentioned SIEM systems.

Next, tables of descriptive statistics for each of the indicators affecting the risks to the UDEE IS were obtained in the STATISTICA 12.5 package, e.g., see Fig. 6.

The study resulted in mean and confidence interval plots for the variables in each cluster characterizing the users of the UDEE in the context of IS compliance and the impact of their work style on IS risks, see Figure 7.

30 листопада 2023, 6:07:06 AM	The user with id '45507' viewed the course with id '4571'.	web	91.123.150.72
30 листопада 2023, 12:50:05 AM	The user with id '82338' viewed the course with id '4571'.	web	46.219.248.141
30 листопада 2023, 12:47:39 AM	The user with id '82338' viewed the user report in the gradebook.	web	46.219.248.141
30 листопада 2023, 12:47:37 AM	The user with id '82338' viewed the course with id '4571'.	web	46.219.248.141
29 листопада 2023, 11:57:38 PM	The user with id '82338' viewed the 'quiz' activity with course module id '454578'.	web	46.219.248.141
29 листопада 2023, 11:57:35 PM	The user with id '82338' viewed the 'quiz' activity with course module id '454578'.	web	46.219.248.141
29 листопада 2023, 11:57:32 PM	The user with id '82338' viewed the 'quiz' activity with course module id '454578'.	web	46.219.248.141
29 листопада 2023, 11:57:08 PM	The user with id '82338' has viewed the submission status page for the assignment with course module id '454580'.	web	46.219.248.141
29 листопада 2023, 11:57:08 PM	The user with id '82338' viewed the 'assign' activity with course module id '454580'.	web	46.219.248.141
29 листопада 2023, 11:57:07 PM	The user with id '82338' created a file submission and uploaded '1' file/s in the assignment with course module id '454580'.	web	46.219.248.141
29 листопада 2023, 11:57:07 PM	The user with id '82338' has uploaded a file to the submission with id '9307987' in the assignment activity with course module id '454580'.	web	46.219.248.141
29 листопада 2023, 11:57:07 PM	The user with id '82338' viewed the 'assign' activity with course module id '454580'.	web	46.219.248.141

Fig. 5. Fragment of LMS Moodle

Variable	Descriptive Statistics for Cluster Cluster contains 12 cases		
	Mean	Standard Deviation	Vari
Year_ study	1,33333	0,492366	0
Pers_ resp	19,75000	6,397798	40
Aver_ perf	68,16666	4,344973	18
Comp_ ass	3,00000	0,852803	0

Fig. 6. Example of descriptive statistics for the variables under consideration

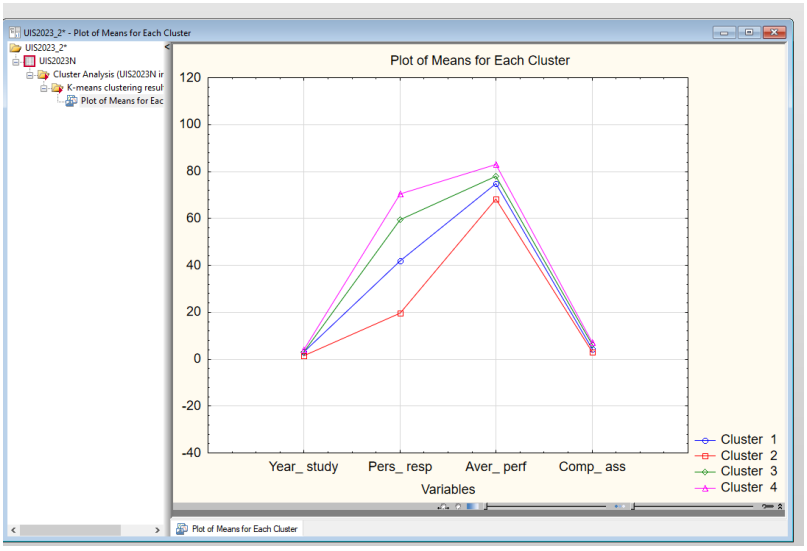

Fig. 7. Results of cluster analysis

VI. Discussion of results

Having compared the k-means method and hierarchical classification, it should be noted that the undoubted advantage of the first method is the ability to work with primary data. This will allow, for example, university IS specialists to process rather large amounts of data, which can be imported from LMS in *.csv format.

In the context of a large information set of users of large educational institutions, this is an undoubted advantage. Moreover, the k-means method can compensate for the consequences of poor quality initial partitioning of the original data set. As shown by the results of the study, analyzing the digital footprints of users of UDEE and analyzing these data on the basis of data mining methods, in particular, cluster analysis is a complex task. Note that in the context of the overall task of improving the security of UDEE, clustering can be an important basic stage. To automate such analysis of DT and in the context of the prospects of continuing research in this direction, it seems reasonable to implement a software multi-agent system, which will contribute to the automated solution of many of the tasks mentioned above. For example, such tasks may include: expert assessment of the level of IS knowledge of UDEE users; semantic analysis of messages in UDEE chat rooms; clustering of users by the degree of influence on IS risks, etc.

Conclusion

The paper presents the results of a study of the state of user competencies for different specialties of the university digital educational environment (UDEE) on issues related to information security (IS). The study was conducted for two large universities in Ukraine and Kazakhstan. The methods of cluster analysis and analysis of digital (electronic) traces (DT) of the users of the UDEE are used. Based on the analysis of the DTs of different groups of registered users in the UDEE, the behavior of six types of users was investigated. These types of users for the specialties considered in the test set, as a result of applying hierarchical classification and k-means method, were divided into appropriate clusters according to the criteria affecting the information security risks of the UDEE. For each cluster, the probability of occurrence of high IS risk incidents can be determined by the UDEE IS expert and, accordingly, measures can be taken to address the causes of such incidents and recommendations for users can be generated.

References

[1] Oliveira, P. C. D., Cunha, C. J. C. D. A., & Nakayama, M. K. (2016). Learning Management Systems (LMS) and e-learning management: an integrative review and research agenda. JISTEM-Journal of Information Systems and Technology Management, 13, 157-180. https://doi.org/10.4301/S1807-17752016000200001

[2] Aldiab, A., Chowdhury, H., Kootsookos, A., Alam, F., & Allhibi, H. (2019). Utilization of Learning Management Systems (LMSs) in higher education system: A case review for Saudi Arabia. Energy Procedia, 160, 731-737. https://doi.org/10.1016/j.egypro.2019.02.186

[3] Azcona, D., Hsiao, IH. & Smeaton, A.F. Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. User Model User-Adap Inter 29, 759–788 (2019). https://doi.org/10.1007/s11257-019-09234-7

[4] Nai, R., Sulis, E., Marengo, E., Vinai, M., Capecchi, S. (2023). Process Mining on Students' Web Learning Traces: A Case Study with an Ethnographic Analysis. In: Viberg, O., Jivet, I., Muñoz-Merino, P., Perifanou, M., Papathoma, T. (eds) Responsive and Sustainable Educational Futures. EC-TEL 2023. Lecture Notes in Computer Science, vol 14200. Springer, Cham. https://doi.org/10.1007/978-3-031-42682-7_48

[5] Mohssine, B., Mohammed, A., Abdelwahed, N., & Mohammed, T. (2021). Adaptive help system based on learners 'digital traces' and learning styles. International Journal of Emerging Technologies in Learning (iJET), 16(10), 288-294. https://doi.org/10.3991/ijet.v16i10.19839

[6] Ye, D., & Pennisi, S. (2022). Using trace data to enhance students' self-regulation: A learning analytics perspective. The Internet and Higher Education, 54, 100855. https://doi.org/10.1016/j.iheduc.2022.100855

[7] Noskova, T., Pavlova, T., & Yakovleva, O. (2018). Study of students' educational activity strategies in the social media environment. E-learning and Smart Learning Environment for the Preparation of New Generation Specialists, 10, 113-125.

[8] N. Kadoić and D. Oreški, "Analysis of student behavior and success based on logs in Moodle," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018, pp. 0654-0659, https://doi.org/10.23919/MIPRO.2018.8400123

[9] Mogus, A. M., Djurdjevic, I., & Suvak, N. (2012). The impact of student activity in a virtual learning environment on their final mark. Active Learning in Higher Education, 13(3), 177-189. https://doi.org/10.1177/1469787412452985

[10] Stiller, K., & Bachmaier, R. (2018, June). Identifying learner types in distance training by using study times. In EDEN Conference Proceedings (No. 1, pp. 78-86). https://doi.org/10.38069/edenconf-2018-ac-0012

[11] Ahmed, A. I., Alharthe, R. M., & Alfereej, M. M. (2023). Organizational committees and their role in enhancing intellectual security: a case study on female students of the Bachelor of Information Science program-College of Arts-Imam Abdul Rahman bin Faisal University. Library Philosophy and Practice, 1-26.

[12] Cheung, S.K.S. (2014). Information Security Management for Higher Education Institutions. In: Pan, JS., Snasel, V., Corchado, E., Abraham, A., Wang, SL. (eds) Intelligent Data analysis and its Applications, Volume I. Advances in Intelligent Systems and Computing, vol 297. Springer, Cham. https://doi.org/10.1007/978-3-319-07776-5_2

[13] Al Quhtani, M. (2017). Data mining usage in corporate information security: Intrusion detection applications. Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy, 8(1), 51-59. https://doi.org/10.1515/bsrj-2017-0005

[14] Salem, I. E., Mijwil, M. M., Abdulqader, A. W., Ismaeel, M. M., Alkhazraji, A., & Alaabdin, A. M. Z. (2022). Introduction to The Data Mining Techniques in Cybersecurity. Mesopotamian journal of cybersecurity, 2022, 28-37. https://doi.org/10.58496/MJBD/2023/007

[15] Kong, J., Yang, C., Wang, J., Wang, X., Zuo, M., Jin, X., & Lin, S. (2021). Deep-stacking network approach by multisource data mining for hazardous risk identification in IoT-based intelligent food management systems. Computational Intelligence and Neuroscience, Volume 202, Article ID 1194565. https://doi.org/10.1155/2021/1194565

[16] Mathew, A. (2023). The Power of Cybersecurity Data Science in Protecting Digital Footprints. Cognizance Journal of Multidisciplinary Studies, Vol.3, Issue.2, February 2023, pg. 1-4. https://doi.org/10.47760/cognizance.2023.v03i02.001

[17] Muhammad, S. S., Dey, B. L., & Weerakkody, V. (2018). Analysis of factors that influence customers' willingness to leave big data digital footprints on social media: A systematic review of literature. Information Systems Frontiers, 20, 559-576. https://doi.org/10.1007/s10796-017-9802-y

[18] Cheng, F. C., & Wang, Y. S. (2018). The do not track mechanism for digital footprint privacy protection in marketing applications. Journal of Business Economics and Management, 19(2), 253-267. https://doi.org/10.3846/jbem.2018.5200

[19] Bollé, T., & Casey, E. (2018). Using computed similarity of distinctive digital traces to evaluate non-obvious links and repetitions in cyber-investigations. Digital Investigation, 24, S2-S9. https://doi.org/10.1016/j.diin.2018.01.002

[20] Sarini, Marcello, Rossana Actis Grosso, Maria Elena Magrin, Silvia Mari, Nadia Olivero, Giulia Paganin, and Silvia Simbula. 2022. "A Cluster Analysis of the Acceptance of a Contact Tracing App—The Identification of Profiles for the Italian Immuni Contact Tracing App" *Healthcare* 10, no. 5: 888. https://doi.org/10.3390/healthcare10050888

[21] Romesburg, C. (2004). Cluster analysis for researchers. Lulu. com. 334 pp.