CTE Workshop Proceedings, 2024, Vol. 11, pp. 273-302 https://doi.org/10.55056/cte.664

Empirical comparison of clustering and classification methods for detecting Internet addiction

Oksana V. Klochko[1], *Vasyl M.* Fedorets[2] and *Vitalii I.* Klochko[3]

[1]*Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University, 32 Ostrozhskogo Str., Vinnytsia, 21100, Ukraine*
[2]*Vinnytsia Academy of Continuing Education, 13 Hrushevskoho Str., Vinnytsia, 21050, Ukraine*
[3]*Vinnytsia National Technical University, 95 Khmelnytsky Hwy, Vinnytsia, 21021, Ukraine*

Abstract. Machine learning methods for clustering and classification are widely used in various domains. However, their performance and applicability may depend on the characteristics of the data and the problem. In this paper, we present an empirical comparison of several clustering and classification methods using WEKA, a free software for machine learning. We apply these methods to the data collected from surveys of students from different majors, aiming to detect the signs of Internet addiction (IA), a behavioural disorder caused by excessive Internet use. We use Expectation Maximization, Farthest First and K-Means for clustering, and AdaBoost, Bagging, Random Forest and Vote for classification. We evaluate the methods based on their accuracy, complexity and interpretability. We also describe the models developed by these methods and discuss their implications for identifying the respondents with IA symptoms and risk groups. The results show that these methods can be effectively used for clustering and classifying IA-related data. However, they have different strengths and limitations when choosing the best method for a specific task.

Keywords: machine learning, clustering, classification, Internet addiction, WEKA

1. Introduction

The empirical analysis of problem-solving methods using machine learning, particularly in sectors such as education, healthcare, and life safety, is a burgeoning area of research [3, 4, 10, 22–24, 47–49]. This paper delves into the exploration of such methods, explicitly focusing on clustering and classification problems [13, 36].

Clustering methods are statistical tools for data analysis that facilitate grouping data samples into clusters or classes based on attribute values. Each group possesses distinct characteristics. The crux of this study is to employ various clustering methods for an empirical comparative analysis to ascertain the most optimal data grouping for a specific problem.

Machine learning categorises clustering problems under unsupervised learning. Many machine learning algorithms exist for clustering such as Expectation Maximization, Farthest First, K-Means, K-Medians, Hierarchical Clustering, and more. However, their applicability varies

klochkoob@gmail.com (O. V. Klochko); bruney333@yahoo.com (V. M. Fedorets); vi.klochko.7@gmail.com (V. I. Klochko)
https://sites.google.com/view/klochko-oksana-v (O. V. Klochko); https://scholar.google.com/citations?user=sfjR5w0AAAAJ (V. M. Fedorets); http://klochkovitaliy.vk.vntu.edu.ua/ (V. I. Klochko)
0000-0002-6505-9455 (O. V. Klochko); 0000-0001-9936-3458 (V. M. Fedorets); 0000-0002-9415-4451 (V. I. Klochko)

depending on the specific problem at hand. These algorithms differ in terms of cluster model type, algorithm model type, cluster nesting hierarchy, and implementation method based on the dataset. Consequently, there are specific requirements for dataset parameters.

Classification tasks form a separate group of machine learning tasks under supervised learning. However, some classification algorithms combine supervised and unsupervised learning (e.g., Learning Vector Quantization). These algorithms are built on a preset finite number of objects divided into groups and can classify an arbitrary object if its group membership is unknown. Numerous machine learning algorithms exist for classification tasks, such as Support Vector Machines (SVMs), Logistic Regression, K-nearest neighbour, Linear Discriminant Analysis (LDA), etc. Ensemble methods are currently gaining popularity due to their increased accuracy.

Several machine learning software products, including TensorFlow, WEKA, MATLAB, MXNet, Torch, PyTorch, and Microsoft Azure Machine Learning Studio, are popular.

In this study, we employ the WEKA (Waikato Environment for Knowledge Analysis) free machine learning software [43], which provides direct access to a library of implemented algorithms written in Java.

A review of contemporary studies and publications reveals that the issue of analysing and selecting an optimal machine learning method for processing a specific dataset is a hot topic in scientific circles. Many of these studies focus on applying machine learning methods in education, healthcare and life safety sectors.

Sentiment analysis is gaining traction in the healthcare and education sectors. For instance, Pacol and Palaoag [34] conducted sentiment analysis on students' textual feedback regarding professors' performance using Machine Learning Techniques. The study found that the Random Forest algorithm was most effective for sentiment classification; it outperformed base models of Support Vector Machines, Naive Bayes, and Logistic Regression algorithms and their ensembles [34]. Klochko et al. [26] utilised machine learning clustering algorithms to analyse common mistakes made by students and to tailor educational content to specific student groups for flipped learning using a virtual learning environment. The analysis was conducted by comparing clusters, defined by student learning outcomes, using Canopy, Expectation Maximization, and Farthest First algorithms.

Souri et al. [37] proposed a model based on Internet of Things technologies for monitoring student health indicators to detect biological and behavioural changes. In conjunction with the Support Vector Machine (SVM) algorithm, the developed model achieves a high accuracy of 99.1% [37].

Hussain et al. [18] explored the application of machine learning classification methods to facilitate independent daily living for individuals with Alzheimer's disease. The study analysed data recorded by various equipment to identify changes in a person's behaviour relevant to daily life and social interaction. The paper compares the efficiency of five machine learning classification techniques for recognising a person's activity (and psychological status). Experimental findings indicate that these approaches yield superior results in determining a person's activity and psychological and behavioural characteristics compared to traditional methodologies.

Krämer, Schreyögg and Busse [28] investigated the speed and efficiency of medical aid provision using Hospital ER databases. The authors developed a model based on preliminary patient diagnosis data by applying the Random Forest algorithm. Using supervised machine

learning methods and model training based on specialist doctor opinions resulted in high forecasting accuracy (96%) and an area under the receiver operating curve greater than 0.99.

Subasi, Kevric and Abdullah Canbaz [38] developed a hybrid model for detecting epileptic seizures using the Genetic Algorithm (GA) and Particle Swarm Optimisation (PSO) to determine the optimal parameters for applying the Support Vector Machine (SVM) algorithm. The proposed hybrid algorithm demonstrated a data set classification accuracy of up to 99.38%.

The emergence and proliferation of the Internet have brought numerous benefits, but concurrently, disorders related to pathological Internet use are becoming a significant social and psychological problem. The detection and prevention of pathologies and premorbid conditions (states before the disease) caused by inappropriate Internet use is a pressing psychological, sociocultural, and educational issue. Internet Addiction (IA) cases were first reported in 1995 and have since garnered considerable attention. The subject has been the focus of research by Derhach [11], Yuryeva and Bolbot [46] among others.

Internet Addiction Disorder (IAD), also known as Pathological Internet Use (PIU), was first proposed by Ivan K. Goldberg in 1995. He characterised net addiction as a specific pathology with a broad spectrum of behavioural and impulse control disorders (lack of control, absence of voluntary regulation) [1]. In 1996, Goldberg first attempted to identify groups of behavioural and psychological signs and symptoms of IA [42]. In 1998, Young [44, 45] defined IAD as an impulsive-compulsive disorder with specific signs or addictions.

Despite the increasing relevance of IA, there needs to be more scientific literature exploring this issue using machine learning methods. A few notable studies include Di et al. [12], which demonstrated the utility of machine learning methods for detecting and forecasting the risk of IA using the Support Vector Machine algorithm on a dataset from a survey conducted among 2,397 Chinese students. Hsieh et al. [16] proposed using the EMBAR-protected system of web services based on ensemble classification methods and case-based reasoning to study user IA and prevent its development at initial stages. Ji, Chen and Hsiao [19] is researching creating an IA detector that works in real-time mode using an adapted system of continuous real-coded variables (XCSR). Suma, Nataraja and Sharma [40] explored the possibilities for predicting IA based on a set of predictor variables using the Random Forest algorithm.

Given the above problem statement and considering the limited research on applying machine learning methods to diagnose IA, our research aims to identify the use fields and conduct an empirical comparison of ensemble classification and clustering methods of machine learning in studying IA.

2. Selection of methods and diagnostics

The study of pupils' IA disorder had two stages. The first study was conducted in 2019. Its purpose was to determine the possible fields of use as well as an empirical comparison of clustering methods of machine learning for studying students' IA disorders. During the second stage, in 2019–2021, the authors studied possible fields of use and an empirical comparison of ensemble classification methods of machine learning for studying IA disorders of students.

In the first stage, data regarding the spread and severity of IA among students majoring in Computer Sciences were received from an online survey, which used a questionnaire drafted

with the help of Google Forms. Two hundred sixty-two students majoring in computer science and coming from different regions of Ukraine participated in the experimental study. The data set is presented in the ARFF format and consists of 8 attributes (figure 1) [25]. The data set contains the fields described in table 1.

```
@relation answer_IA

@attribute age numeric
@attribute sex {female,male}
@attribute 3 {no,undefined,yes}
@attribute 4 {no,undefined,yes}
@attribute 5 {no,undefined,yes}
@attribute 6 {no,undefined,yes}
@attribute 7 {no,undefined,yes}
@attribute 8 {no,undefined,yes}

@data
18,male,yes,no,no,no,no,yes
28,male,undefined,no,no,no,no,yes
20,female,yes,yes,yes,no,no,no
22,male,yes,no,no,no,no,no
...
```

Figure 1: Data set on the state of IA among students majoring in Computer Sciences, presented in the ARFF format.

Cluster analysis is one of the tasks of database mining. Cluster analysis is a set of methods of multidimensional observations or object classification based on defining the distance between the objects and their subsequent grouping (into clusters, taxons, and classes). The selection of a concrete cluster analysis method depends on the purpose of classification [27]. At the same time, one does not need a priori information about the statistical population. This approach is based on the following presuppositions: objects with a certain number of similar (different) features group in one segment (cluster). The level of similarity (difference) between the objects that belong to one segment (cluster) must be higher than the level of their similarity with the objects that belong to other segments [27].

Let us look at one of the cluster analysis algorithms [27].

Output matrix:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix}.$$

Let us move to the matrix of standardised Z values with elements:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j};$$

Table 1
Data structure on the state of IA among students majoring in Computer Sciences.

Attributes	Contents/Questions	Type	Statistics
age	Age of the student	Numeric	Minimum 16 Maximum 59 Mean 19.756 StdDev 6.806
sex	Student's sex	Nominal	Female 199 Male 63
3	Can't imagine my life without the Internet	Nominal	yes 184 undefined 39 no 39
4	When I cannot use the Internet I feel anxiety, irritation	Nominal	yes 81 undefined 134 no 47
5	I like "surfing" the Net without a clearly defined purpose	Nominal	yes 121 undefined 112 no 29
6	I can give up food, sleep, and going to classes if I have a chance to use the Internet for free	Nominal	yes 248 undefined 7 no 7
7	I prefer meeting new people over the Internet rather than in real life	Nominal	yes 185 undefined 37 no 40
8	I often feel that I have not spent enough time playing computer games over the Internet, I constantly wish to play longer	Nominal	yes 178 undefined 61 no 23

where $j = 1, 2, \cdots, n$ — index number, $i = 1, 2, \cdots, m$ – observation number;

$$\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij};$$

$$s_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \overline{x}_j)^2} = \sqrt{\overline{(x_{ij}^2)} - (\overline{x}_j)^2}.$$

There are several ways to define the distance between two observations z_i and z_v: weighted Euclidean distance, which is determined by the formula

$$\rho_{BE}(z_i, z_v) = \sqrt{\sum_{l=1}^{n} w_l (z_{il} - z_{vl})^2},$$

where w_l is the "weight" of index; $0 < w_l \leq 1$; if $w_l = 1$ for all $l = 1, 2, \ldots, n$, then we get the usual Euclidean distance

$$\rho_{BE}(z_i, z_v) = \sqrt{\sum_{l=1}^{n}(z_{il} - z_{vl})^2}.$$

Hamming distance:

$$\rho_{BH}(z_i, z_v) = \sum_{l=1}^{n}|z_{il} - z_{vl}|.$$

In most cases, this distance measuring method gives the same result as the usual Euclidean distance, but in this case, the influence of significant non-systemic differences (runouts) decreases.

Chebyshev distance:

$$\rho_{BCH}(z_i, z_v) = max_{1 \leq l \leq n}|z_{il} - z_{vl}|.$$

It is best to apply this distance to determine the differences between the two objects using only one dimension.

Mahalanobis distance:

$$\rho_{BM}(z_i, z_v) = \sqrt{(z_i - z_v)^T S^{-1}(z_i - z_v)},$$

where S is the covariance matrix; this distance measurement gives good results when applied to a concrete data group. However, it only works well if the covariance matrix is calculated for the whole data set.

Distance between peaks:

$$\rho_{BL}(z_i, z_v) = \frac{1}{n}\sum_{l=1}^{n}\frac{|z_{il} - z_{vl}|}{z_{il} + z_{vl}},$$

presupposes the independence of random variables, which indicates the distance in the orthogonal space.

It is best to choose from the distance measures described above after considering the structure and characteristics of the data sample.

Let us present the received measurements in the form of a distance matrix:

$$R = \begin{pmatrix} 0 & \rho_{12} & \rho_{13} & \ldots & \rho_{1m} \\ \rho_{21} & 0 & \rho_{23} & \ldots & \rho_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{i1} & \rho_{i2} & \rho_{i3} & \ldots & \rho_{im} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{m1} & \rho_{m2} & \rho_{m3} & \ldots & 0 \end{pmatrix}.$$

As the R matrix is symmetric, i.e. $\rho_{iv} = \rho_{vi}$, we may confine ourselves to off-diagonal matrix elements. We can implement the cluster analysis agglomerative hierarchic procedure using

the distance matrix. Distances between clusters are determined as the closest or the farthest ones. In the first case, the distance between the clusters is between the closest elements of these clusters; in the second case, it is between the two farthermost locations. The principle of agglomerative hierarchic procedures lies in a consequent grouping of elements, starting from the ones closest to each other and those farther and farther apart. During the first step of the algorithm, every observation z_i $(i = 1, 2, ..., m)$ is viewed as a separate cluster. Then, during every next step of the algorithm, the two closest located clusters are grouped, and then once again, the distance matrix is built, but its dimension decreases by one. The algorithm stops its work when all the observations are grouped into clusters.

Let us look at the algorithms we used while clustering the data set regarding the state of IA disorder among students majoring in Computer Sciences:

EM (Expectation Maximization):

Determines the probability distribution for every object, which indicates its belongingness to each cluster. EM methods [20]: Maximum Likelihood Estimation (MLE) or Maximum a Posteriori (MAP). Description of the algorithm is shown in figure 2 [20]: at the E-stage (expectation), we calculate the estimated likelihood; at the M-stage (Maximization), we calculate the maximum likelihood estimation, increasing the expected likelihood, calculated at the E-stage; its value is used for the E-stage at the next iteration. The algorithm is repeated until its convergence.

0. **Initialization**: Get an initial estimate for parameters θ^0 (e.g. all the μ_k, σ_k^2, and π variables). In many cases, this can just be a random initialization.
1. **Expectation Step**: Assume the parameters (θ^{t-1}) from the previous step are fixed, compute the expected values of the latent variables (or more often a *function* of the expected values of the latent variables).
2. **Maximization Step**: Given the values you computed in the last step (essentially known values for the latent variables), estimate new values for θ^t that maximize a variant of the likelihood function.
3. **Exit Condition**: If likelihood of the observations have not changed much, exit; otherwise, go back to Step 1.

Figure 2: Description of how the algorithm EM works from 10,000 feet [20].

K-Means algorithm:

It aims to partition n observations into k clusters so that each observation belongs to the cluster with the nearest mean value. The shortest distance between the observations and the nearest mean value may be calculated by minimising the sum of squares of the distances [32] (figure 3).

Farthest First algorithm:

This is a modification of a K-Means algorithm, in which the initial selection of centroids is two and higher. Centroids are determined following the remoteness principle, i.e. the point farthest from the rest is selected first. Figure 4 [9] describes the Farthest First algorithm.

The second stage of the survey on the situation with IA among students of various specialities was conducted with the help of Google Forms. Three hundred sixty-three students from different regions of Ukraine took part in the survey. The data set is presented in the ARFF format and

Require: c – number of clusters
Initialization: Randomly select c points
that will be cluster centroids for first iteration.
repeat
Assign each observation from the to the cluster
with the nearest centroid. Recalculate cluster centroids
taking into consideration the current observation distribution.
until Until the structure stabilizes or the condition for stopping
the algorithm is fulfilled (e.g. maximal number of iterations)

Figure 3: K-Means algorithm [32].

Input: n data points with a distance metric $d(\cdot, \cdot)$.

Pick a point and label it 1.

For $i = 2, 3, \ldots, n$
Find the point furthest from $\{1, 2, \ldots, i-1\}$ and label it i.
Let $\pi(i) = \arg\min_{j<i} d(i, j)$.
Let $R_i = d(i, \pi(i))$.

Figure 4: Farthest-first traversal of a data set [9]. Take the distance from a point x to a set S to be $d(x, S) = min_{y \in S} d(x, y)$ [9].

consists of 9 attributes (figure 5). The data set contains the fields described in table 2.

Cluster analysis is one of the tasks of database mining. Cluster analysis is a set of methods of multidimensional observations or object classification based on defining the distance between the objects and their subsequent grouping (into clusters, taxons, and classes). The selection of a concrete cluster analysis method depends on the purpose of classification [27]. At the same time, one does not need a priori information about the statistical population. This approach is based on the following presuppositions: objects that have a certain number of similar (different) features grouped in one segment (cluster). The level of similarity (difference) between the objects that belong to one segment (cluster) must be higher than the level of their similarity with the objects that belong to other segments [27].

In order to analyse the IA phenomenon, we divide the respondents into three groups (Significant Risk (SR), Insignificant Risk (IR), and No Risk (NR)). The division is based on the integrative use of qualitative and quantitative characteristics of the IA phenomenon. The SR group is formed based on detecting and analysing those IA features, which signify qualitative changes in the psychological status of a personality. The selection of such features is based on a traditional understanding that in-depth psychic changes related to the formation of addictive behaviour concern, primarily vital [5] and existential "foundations" of a personality. Such vital and existential (ontological) "foundations" are relatively stable and are subject to "external" transformation if the influence is significant and long-lasting. The changes concern the existential dimension [14], vital resources [5], intentions, attitudes, and behavioural stereotypes aimed at survival and life preservation. They are linked with the vital "foundations" of life itself.

```
@relation answer_363_IA

@attribute age numeric
@attribute sex {female,male}
@attribute 3 {no,undefined,yes}
@attribute 4 {no,undefined,yes}
@attribute 5 {no,undefined,yes}
@attribute 6 {no,undefined,yes}
@attribute 7 {no,undefined,yes}
@attribute 8 {no,undefined,yes}
@attribute IA {nr,ir,sr}

@data
19,female,yes,yes,yes,no,no,no,sr
24,male,yes,yes,no,no,undefined,undefined,sr
26,female,no,no,no,no,no,no,nr
19,male,yes,no,no,no,yes,yes,ir
...
```

Figure 5: Data set on the state of IA of students, presented in the ARFF format.

The questions which reflect the above-stated life "foundations" or vital resources are (table 2): "I can give up food, sleep, going to classes, if I have a chance to use the Internet for free" (1^{st} SR) and "When I cannot use the Internet, I feel anxiety, irritation" (2^{nd} SR). The 1^{st} SR seems more important as food and sleep are system organising and basic vital needs. The ability to give them up indicates not only the "total" in-depth and comprehensive change of the hierarchy of vital needs, values and senses [14, 31], but also the deformation of a very vital "foundation" of a personality. The stated issues of food and sleep indirectly reflect a person's existential problems. This is caused by the fact that these issues concern the existential problem of "life and death" and the "I am" existential phenomenon. That is why, while IA is being formed, the existential problem is also being developed, which is temporarily and compensatory solved with a "potential possibility of Internet access".

In the first (1^{st} SR) question, the "can give up classes" part is an important social and personality-oriented aspect. If the answer is positive, the content embedded in the above fragment is ignored and desemantized. This discloses the presence of desemantization and depreciation of the possible socio-economically "settled" future and a conscious self-limitation in the field of self-actualisation in studying and professional activity. Another relevant point is ignoring communication, social ties, possibilities for self-improvement and "construction" of self in the educational discourse. The stated needs and aims are partially changed and substituted by the Internet. At the same time, the "real" reality is replaced, deactualized and desemantized. The second (2^{nd} SR) question reflects the presence of neurotic anxiety, which is a manifestation of "exhaustion" and "overstrain" of the nervous system as well as of certain changes in the system of emotions and volition. Moreover, in this case, the problem of sense formation and understanding and the corresponding changes in the value-conceptual field occur. This is what

Table 2
Data structure on the state of IA of students.

Attributes	Contents/Questions	Type	Statistics
age	Age of the student	Numeric	Minimum 15 Maximum 59 Mean 20.306 StdDev 7.238
sex	Student's sex	Nominal	Female 260 Male 103
3	Can't imagine my life without the Internet	Nominal	yes 245 undefined 53 no 65
4	When I cannot use the Internet I feel anxiety, irritation	Nominal	yes 110 undefined 194 no 59
5	I like "surfing" the Net without a clearly defined purpose	Nominal	yes 169 undefined 156 no 38
6	I can give up food, sleep, and going to classes if I have a chance to use the Internet for free	Nominal	yes 343 undefined 8 no 12
7	I prefer meeting new people over the Internet rather than in real life	Nominal	yes 263 undefined 48 no 520
8	I often feel that I have not spent enough time playing computer games over the Internet, I constantly wish to play longer	Nominal	yes 243 undefined 86 no 343
IA	IA disorder	Nominal	nr 113 ir 217 sr 33

V. Frankl [14] described and logoneurosis – the loss and/or absence of sublime and vital senses. The presence of a neurotic aspect in the form of neurotic anxiety is particularised through actualisation in the 2^{nd} SR question of the irritation phenomenon ("… feel … irritation").

In general, the 2^{nd} SR question supplements, particularises and "strengthens" the deficit and "narrowing" of vitality, life creativity, nature corresponding existence, and healthy life preservation instinct (food, sleep, communication). The stated vital resources and life creativity, actualised and problematised in the 1^{st} SR and 2^{nd} SR question, act as a complex diagnostic subsystem. It aims to detect systemic, comprehensive, stable, in-depth, vital, personal-psychological problems (i.e. disorders). The stated problems may develop and together transform into AI.

The 1^{st} SR and 2^{nd} SR questions, which reflect the vitality of a person and his/her existence, disclose the qualitative difference of the SR group from IR and NR groups, which do not have the

stated peculiarities. Thus, with a certain degree of certainty, the SR group may be represented as well as diagnosed with a relatively limited number of questions (1^{st} SR and 2^{nd} SR). The questions, which represent the SR group, indicate that the Internet has "penetrated" deep into the consciousness and the core of a personality, in his/her vitality, into human existence (figure 6).

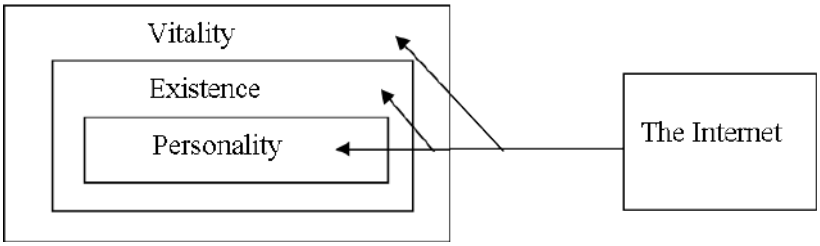

Figure 6: "Penetration" of the Internet into the core of a personality, into vitality, into existence, which illustrates the psychological mechanisms of risk formation in the SR group.

Thus, the SR group is qualitatively different from IR and NR. The SR group represents either a considerable risk of IA development or a transitive (premorbid) state or even the presence of the actual IA pathology. In contrast, the IR and NR groups indicate a greater or lesser possibility of its development.

The IR group is diagnosed by questions 3, 5, 7 and 8 of table 2, which reflect "weak" signs. The signs reflected in these questions are not attributive or essential. Accordingly, they do not represent IA's in-depth personality-psychological, vital and existential aspects. That is why these questions, by summing up a certain number of them (in this case, three), may form a certain degree of probability for having IA risks. These questions get a certain level of "consistency" when there is a certain number of them, in this case, not less than three.

Thus, the IR group is characterised (diagnosed) by the presence of three questions out of questions 3 (1^{st} IR), 5 (2^{nd} IR), 7 (3^{rd} IR) and 8 (4^{th} IR) of table 2. The 1^{st} IR question discloses the contemporary reality of professional activity and communication, in which the Internet component is relevant, systemic, environmental and significant. Thus, taking into consideration current systemic Internet-oriented socio-technological and technological contexts, this question does not reveal the totality and explicitness of personality-psychological changes. It primarily indicates a considerably high level of significance and even value of the Internet in a person's life. The 2^{nd} IR question characterises the peculiarity of the contemporary Internet culture. It reflects the fact that the person has an actualised orientation-searching reflex and a corresponding search and cognitive behaviour, and not just that possibility of IA development. While disclosing the peculiarities of modern-day Internet communication, the 3^{rd} IR question also partially characterises the problem of a personality's insufficiently developed communicative competence and communicative culture, which, in turn, is effectively compensated with Internet communication.

As for the 4^{th} IR, the significant question is the one which indicates an integrative manifestation of the activity (leading aspect), cognitive, value-conceptual, creative and communicative dimensions of the psychic reality. The positive answer partially indicates the presence of insufficient risks. At the same time, in its essence, a person is a creature that plays – Homo ludens

[17]. What is essential is that in early childhood, a game is a specific integrative and integrating form of activity and the essence of human existence. The state gaming essence of a person can manifest itself in solving complicated tasks and studying as well as during leisure time. A computer game has a considerable mobilising, emotional, orientation-search potential. A computer game can become addictive due to the "gaming" peculiarities of human nature and as a result of the professionally developed games that take human psychology into account. A computer game addiction thus indicates not only not as much the risks of IA development but rather the presence of the "gaming essence" of Homo Ludens. In addition, a considerable number of teenagers and grown-ups develop computer game addiction as a consequence of the fact that they did not have a chance to realise their gaming potential in early childhood. This often happens due to intense early learning, which competes with gaming activity. The stated principle of competition between different forms of activity is described in Anokhin [2] study on functional systems [39]. At the same time, constant interest in playing computer games poses a specific threat to developing computer games and Internet addictions as it "touches" different psychic spheres.

At the same time, if a person gives at least three positive answers to the 1^{st} IR, 2^{nd} IR, 3^{rd} IR and 4^{th} IR questions, this indicates a particular risk of IA development. This is caused by the fact that each question characterises the influence of the Internet on a certain aspect of a personality: the 1^{st} IR – on the need and value-conceptual aspect, the 2^{nd} – on the cognitive aspect, including the orientation ability; the 3^{rd} – on the communicative aspect, the 4^{th} – on the activity component. Thus, the actualisation of the Internet as a need, value-conceptual, cognitive, communicative and activity phenomenon that is significant for a personality speaks of its particular "spread" and "rootedness" in the stated aspects (spheres) as well as about its corresponding significance and value. A particular "Internet locus" is formed in various personality spheres. Thus, as a result of systemic interiorisation processes, the Internet "integrates" into a person's consciousness, becoming a significant phenomenon (figure 7). At the same time, the stated "integration" is "superficial", reversal, unstable, and such that does not lead to maladaptation or personality-psychological and behavioural changes.

The totality of the "spread" or "expansion" of the Internet on the psychic reality, as a significant phenomenon for several spheres of consciousness, creates certain (but considerably insignificant) risks of IA development. At the same time, the more spheres "contain" the "Internet locus", the higher the risks are, as this means the increase of opportunities for forming the summing up and synergy effects, which lead to qualitative changes.

NR is a group of questions which characterise separate features of Internet influence. The

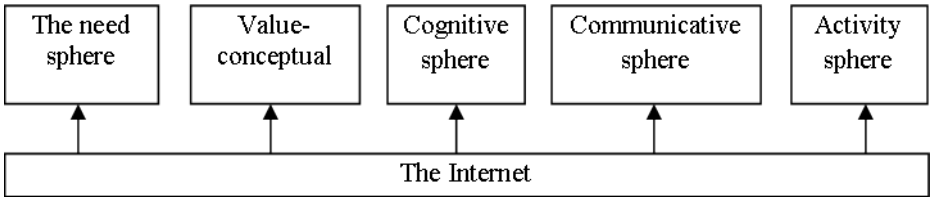

Figure 7: Influence of the Internet on various spheres of human psychic.

personality demonstrates local, superficial, reactive reactions. That is why the risks are almost absent or relatively insignificant for the people in this group.

The classification ensemble methods of machine learning were used to study IA disorder in students. The ensemble methods combine a few algorithms that learn simultaneously and compensate or correct mistakes of one another. Such approaches as stacking, bagging (bootstrap aggregating) and boosting are used while developing ensemble methods. Stacking used the meta-learning approach to best combine a few machine learning models. The algorithm is taught based on the basic-level models. Using these results, the meta-model learns to combine basic model predictions better. Bagging uses multiple teachings of an ensemble of classifiers on random data sets conducted simultaneously but independently from one another. Then, a determined averaging of results is conducted. The results are averaged based on a determined strategy. Boosting carries out a consecutive adaptive algorithm teaching. The following algorithm learns by focusing on the classification mistakes of the first algorithm. The authors used ensemble classification algorithms such as AdaBoost, Bagging, Random Forest, and Vote in this research.

The WEKA machine learning system uses the AdaBoost algorithm.M1 (figure 8) [15, 35]. The AdaBoost meta-algorithm improves the efficiency of basic learning algorithms by building their combination. It uses adaptive boosting, building every next classifier according to the instances badly classified by previous classifiers. Having determined a weak classifier in the cycle, AdaBoost re-assigns the weights, and at every iteration, the weights of incorrectly classified instances increase. By testing classifiers in such a way, the AdaBoost algorithm selects a classifier that better identifies the instances.

Input: sequence of m examples $\langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$
with labels $y_i \in Y = \{1, \ldots, k\}$
weak learning algorithm **WeakLearn**
integer T specifying number of iterations

Initialize $D_1(i) = 1/m$ for all i.

Do for $t = 1, 2, \ldots, T$:

1. Call **WeakLearn**, providing it with the distribution D_t.
2. Get back a hypothesis $h_t : X \to Y$.
3. Calculate the error of h_t: $\epsilon_t = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$.

   If $\epsilon_t > 1/2$, then set $T = t - 1$ and abort loop.
4. Set $\beta_t = \epsilon_t / (1 - \epsilon_t)$.
5. Update distribution D_t:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t & \text{if } h_t(x_i) = y_i \\ 1 & \text{otherwise} \end{cases}$$

where Z_t is a normalization constant (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$h_{fin}(x) = \arg\max_{y \in Y} \sum_{t:h_t(x)=y} \log \frac{1}{\beta_t}.$$

Figure 8: The algorithm AdaBoost.M1 [15].

The Bagging (bootstrap aggregation) meta-algorithm uses compositions of algorithms, each of which learns independently from one another; to determine the final result, the process called voting is being implemented, as a result of which, the mistakes of the classifiers are compensated [6, 41] (figure 9).

For Classification, use a training set **X**, Inducer **I** and the number of bootstrap samples **m** as input. Generate a classifier $\mathbf{F}_{bag}$ as output. $\mathbf{F}_{bag}$ is the bagged prediction, and $\mathbf{F}_1(X)$, $\mathbf{F}_2(X)$,..., $\mathbf{F}_b(X)$ are the predictions from the individual base learners.

1. Create **m** new training sets $\mathbf{X}_i$, from **X** with replacement.
2. Classifier $\mathbf{F}_i$ is built from each set $\mathbf{X}_i$ using **I** to determine the classification of set $\mathbf{X}_i$.
3. Finally classifier $\mathbf{F}_{bag}$ is generated by using the previously created set of classifiers $\mathbf{F}_i$ on the original data set **X**, the classification predicted most often by the sub-classifiers $\mathbf{F}_i$ is the final classification. $\mathbf{F}_{bag} = \mathbf{F}_1(X) + \mathbf{F}_2(X) + \ldots + \mathbf{F}_b(X)$.

Figure 9: The algorithm Bagging [6, 30].

According to Leo Breiman's definition, "a Random Forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1, \ldots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x" (figure 10) [7, 50].

Step 1. Random samples from the given data set are generated.
Step 2. The algorithm constructs a decision tree for each sample, receives the prediction result for each decision tree.
Step 3. Voting for each forecasted result is conducted.

Figure 10: The algorithm Random Forest [7].

The voting classifier combines different classifiers that learn and are assessed simultaneously [21, 29]. The final decision regarding the prediction is taken by a majority vote following two strategies. In hard voting (majority voting), the class label is predicted, which is determined by a majority of votes of every classifier [21, 29]. In soft voting, probability vectors for every predicted class (for all classifiers) are summed up and averaged, and the class with the highest value is selected [21, 29].

3. Results and discussion

To cluster data using the WEKA platform, we will use Weka.clusterers.EM, Weka.clusterers.SimpleKMeans and Weka.clusterers.FarthestFirst algorithms [43].

We check the application of clustering algorithms that can be assigned to two classes of clustering algorithms, i.e. distribution-based (Expectation Maximization) and centroid-based

(K-Means, Farthest First). Such selection is motivated because these algorithms have long been used to cluster different data types in many fields and are considered effective.

Dunn, DB, SD, CDbw and S_Dbw were selected as validity indices for testing [8, 33] (table 3). In the CDbw index, the distance from the point to the multitude set when selecting a cluster element can be calculated differently. In this study, we use the sum of distances of the cluster's already existing "representatives" to each cluster element to calculate this distance. The element on which the maximum was reached was selected as the next "representative" of the cluster.

Table 3
Optimal number of clusters, calculated with the help of quality indices.

Index	Algorithms		
	Expectation Maximization	k-Means	Farthest First
Dunn	3	6	6
DB	3	6	4
SD	3	3	3
CDbw	3	3	3
S_Dbw	3	5	4

If the data set has no cluster structure, then such a situation is not determined with the help of validity metrics. While using K-Means and Farthest First (table 2), the number of clusters for the two algorithms that were selected as optimal by most indices can only be nominally defined as cluster structure. As the work of the Expectation Maximization algorithm is based on determining the probability of evaluating maximum similarity, the indices calculated for this algorithm are more homogenous. The structure, characterised by a small number of clusters that also have to be compact and separable, is determined to be the best. Judging by the results of the clustering evaluation using the validity indices, we may consider that k-Means and Farthest First algorithms are most likely to give worse clustering results than the Expectation Maximization algorithms.

We select training/testing to cluster the data using the percentage split option. As a data set for training (model building), we select 66% of data from the set. As a data set for testing, we select 34% of data from the set. In addition, we select the number of clusters "3" in algorithm settings.

In applying the EM clustering algorithm, according to the built clustering model based on the training data set, three clusters were determined; their characteristics are given in table 4.

Cluster 0 (63% of respondents): The average age of respondents in this cluster is 17. The group consists predominantly of women. The characteristic feature of this group's representatives is that they cannot imagine their life without the Internet. There are variations in the levels of anxiety and irritation if there is no possibility to use the Internet. There are also varying opinions regarding the aimless use of the Internet. As for other attributes, disorders related to IA may be observed in the insignificant number of respondents who belong to this cluster. The behavioural model of the representatives of this cluster demonstrated Internet centration in the psychic reality of a personality, which is accordingly reflected in their activity and behaviour, other life interests, and the significance of everyday activities lose their importance. The stated tendencies are linked to IA.

Table 4
Model and evaluation on test split by EM algorithm.

Attributes	Indications	Clusters		
		0 (0,63) 112.1491	1 (0,13) 24.7781	2 (0,24) 44.0727
age	mean	17.4469	36.2459	19.2906
	std. dev.	1.5994	10.0785	2.243
sex	female	108.8714	16.0638	5.0648
	male	2.2778	7.7143	38.0079
3	no	22.7034	3.1864	10.1102
	undefined	16.0405	6.4026	4.5569
	yes	73.4052	15.1891	29.4057
4	no	54.392	13.8263	27.7817
	undefined	23.6012	5.1903	7.2085
	yes	34.156	5.7615	9.0825
5	no	45.3302	19.3167	26.3531
	undefined	15.1791	2.1415	5.6794
	yes	51.6398	3.32	12.0403
6	no	106.1573	22.7561	41.0866
	undefined	1.0117	1.0098	1.9785
	yes	4.9802	1.0122	1.0076
7	no	81.1224	20.5492	27.3284
	undefined	11.5501	2.168	11.282
	yes	19.4767	2.061	5.4624
8	no	89.4444	19.3333	9.2223
	undefined	7.2533	1.1937	9.553
	yes	15.4514	4.2512	25.2975

Cluster 1 (13% of respondents): For the representatives of this group, the average value of the age attribute is 36, and it varies greatly. This is the oldest age group if compared with other clusters. This group has the largest share of women. Representatives of this group, predominantly, cannot imagine their life without the Internet. Thus, according to the centroid values of the attributes, we may diagnose IA-related Internet centration in the psychic reality of a personality, which is accordingly reflected in their activity and behaviour; other life interests, as well as the significance of everyday activities, lose their importance. There are predominantly no other signs of IA-related disorders.

Cluster 2 (24% of respondents): The probabilistic average of the age attribute among this group's representatives is middle-aged compared to other groups and is 19. Male representatives significantly dominate in this group. Regarding the inability to imagine their life without the Internet, opinions differed; however, most respondents believe they have this addiction. Judging by the values of attributes 4, 5, 6 and 7, most of this group's representatives declare that they

do not have other signs of IA. However, the feeling of the lack of time spent playing computer games over the Internet, confirmed by the vast majority of respondents, is a warning signal that may signify the existence of IA-related disorders. Thus, the characteristic feature of this group is that most of its representatives have IA-related disorders such as Internet centration in the psychic reality of a personality and behavioural impulse control disorders related to online gaming. These people are in the risk group for developing IA-related disorders.

In applying the Farthest First algorithm, according to the built clustering model based on the training data set, three clusters have also been formed; their characteristics are given in table 5.

Table 5
Model and evaluation on test split by Farthest First algorithm.

Attributes	Clusters		
	0	1	2
age	16.0	22.0	20.0
sex	female	male	male
3	yes	undefined	yes
4	undefined	no	yes
5	no	yes	undefined
6	no	no	undefined
7	no	undefined	undefined
8	no	undefined	no

Cluster 0: Contains data instances of the youngest age group, whose age centroid attribute is 16. According to the value of the sex centroid attribute, the group is made up of primarily female data instances. The representatives of this group cannot imagine their lives without the Internet, i.e., there is evident Internet centralisation in the psychic reality of a personality. Respondents cannot determine whether they feel anxiety or irritation if they cannot use the Internet. Judging by other attributes, data instances of this cluster do not have IA-related disorders.

Cluster 1: This cluster contains data instances of an older age group, the age attribute centroid of which is 22. The value of the sex attribute centroid in this cluster is male. A characteristic feature of the cluster is undecidedness regarding the vital need to use the Internet, the prevalence of Internet relations over actual real interactions, and the lack of time spent playing computer games over the Internet (attributes 3, 7, 8 equal undefined). The value of the yes centroid of attribute 5 shows the inclination to use the Internet without a concrete purpose. To give an overall characteristic, this group has signs of IA, i.e. behaviour control disorders related to Internet use.

Cluster 2: By the value of the age attribute centroid, 20, this cluster contains data instances of the middle age group if compared with other clusters. The sex attribute centroid in this cluster is male. The representatives of this cluster cannot imagine their life without the Internet and feel anxiety and irritation when they cannot use the Internet. Their undecidedness characterises them regarding the vital need to use the Internet, giving up other life interests and everyday activities for free Internet use, and the prevalence of online relations of real-life interactions (value of attributes 5, 6, 7 is undefined). Thus, the representatives of this cluster have signs of IA, the priority significance of the Internet and behaviour control disorders related to Internet

use. Compared to other groups, they are in the risk group for developing IA-related disorders.

In applying the K-Means algorithm to the clustering model built based on the training data set, three clusters have also been formed; their characteristics are presented in table 6.

Table 6
Model and evaluation on test split by K-Means algorithm.

Attributes	Clusters		
	0	1	2
age	18.4194	21.8605	20.9552
sex	female	male	female
3	undefined	yes	yes
4	no	no	no
5	no	no	no
6	no	no	no
7	no	no	no
8	no	yes	no

Cluster 0: Contains data instances of the youngest age group, whose age attribute centroid is about 18. According to the sex attribute centroid, mostly female data instances are present in the groups. The representatives of this group need help determining whether they have a vital need to use the Internet. As for other indices, respondents state the absence of signs of IA-related disorders.

Cluster 1: This cluster contains data instances of the older age group, whose age attribute centroid is about 22. The value of the sex attribute centroid in this cluster is male. Characteristic features of data instances that belong to this cluster include the vital need to use the Internet, the lack of time spent playing online computer games, and the systemic need to play longer. The overall characteristic of this cluster is the presence of signs of IA, i.e. behaviour control issues related to Internet use, namely, gaming Internet addiction. If compared with other clusters, they belong to the risk group that may develop IA-related disorders.

Cluster 2: By the value of age attribute centroid, which is about 21 years, compared to other clusters, this cluster contains data instances of medium age group. The sex attribute centroid is female. The representatives of this cluster cannot imagine their life without the Internet. Judging by centroids of other characteristics, respondents of this cluster do not have Internet-related disorders. Thus, the representatives of this cluster have only IA signs associated with the utmost significance of the Internet.

The cluster distribution of test data in applying the three algorithms – the Expectation Maximization, Farthest First and K-Means – using the built training models is presented in table 7. Thus, as can be seen from the table, the algorithms have determined three data groups. Clusters were formed, which included 71:12:7, 67:4:19 and 33:15:42 data instances, respectively. There is a cluster with the largest data instances, a group with the least (exceptions), and a group with several times more data instances than the smallest group.

Figures 11, 12 and 13 present a graphic representation of clusters by age characteristic of data instances, which are built using the training data set and received in the course of implementation of the Expectation Maximization, the Farthest First and the K-Means algorithm

Table 7
Clustered Instances determined using Expectation Maximization, K-Means and Farthest First algorithms.

Attributes	Expectation Maximization		Farthest First algorithm		K-Means	
	Instances	%	Instances	%	Instances	%
0	67	74	71	79	33	37
1	4	4	12	13	15	17
2	19	21	7	8	42	47

respectively. As we can see, the formed clusters differ from each other by the age attribute. For instance, Cluster 0, which contains most data instances, contains instances of respondents of a younger age if formed through the application of the Expectation Maximization algorithm (figure 11). On the other hand, the same cluster received through the implementation of the Farthest First algorithm contains data instances of various age groups (figure 12). Also, a small number of data instances of various age groups is present in Cluster 2, received during the implementing of the K-Means algorithm (figure 13). Cluster 0 and Cluster 2 were formed with the Expectation Maximization algorithm, Cluster 1 and Cluster 2 were formed with the Farthest First algorithm containing homogeneous age groups, and Cluster 0 and Cluster 1 were formed with the K-Means algorithm.

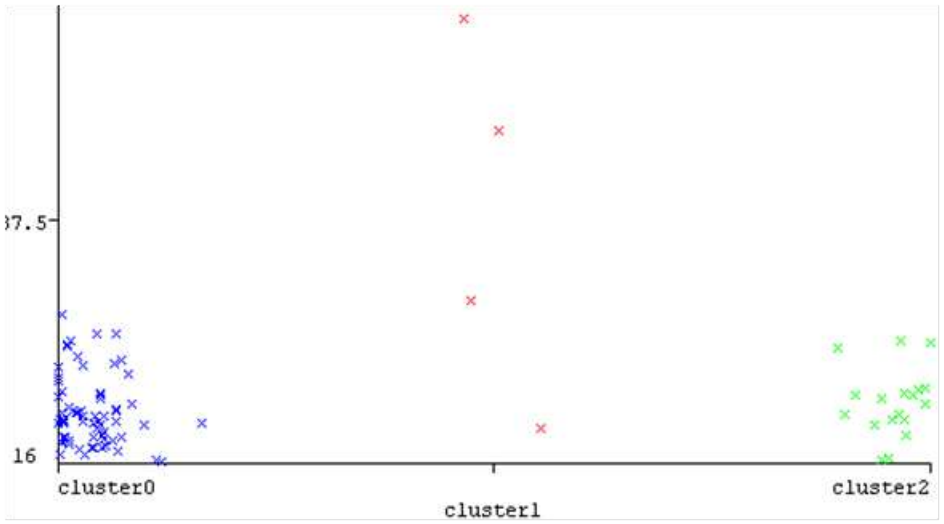

Figure 11: Plot of cluster distribution applying the Expectation Maximization algorithm depending on the age group attribute.

Figures 14, 15 and 16 present a graphic representation by sex attribute of clusters formed through the application of the Expectation Maximization, Farthest First and K-Means algorithm respectively. The analysis of figure 14, which visualises clustering by applying the Expectation Maximization algorithm, shows that Cluster 0 contains only female data instances. Clusters 1 and 2 have date instances of both sexes. Female data instances prevail in Cluster 1 and male ones in Cluster 2. Unlike Clusters formed by the Expectation Maximization algorithm, all the clusters

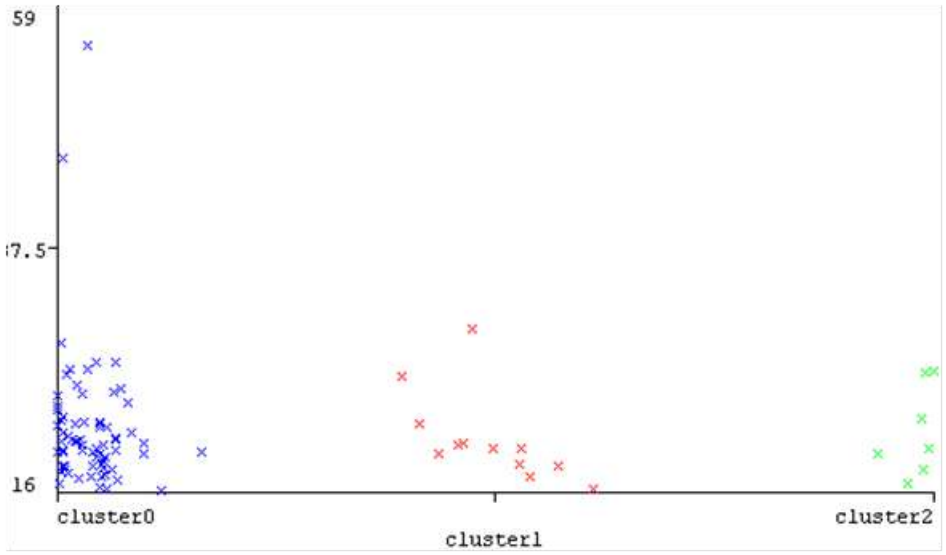

Figure 12: Plot of cluster distribution applying the Farthest First algorithm depending on the age group attribute.

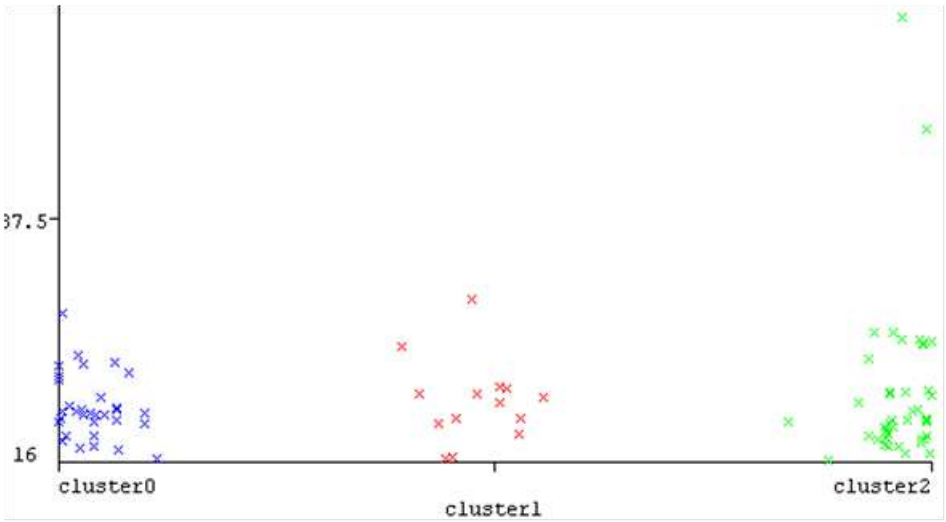

Figure 13: Plot of cluster distribution applying the K-Means algorithm depending on the age group attribute.

formed by the Farthest First algorithm contain data instances of both sex groups (figure 15). Female data instances significantly prevail in Cluster 0. All the clusters built using the K-Means algorithm contain male and female data instances (figure 16).

To classify a data set containing 363 data sets, we break it with the help of random choice into a training set containing 70% (254) data sets and a test set containing 30% (109) data sets.

To classify using the WEKA machine learning system, we create classification models based on the training set with the help of AdaBoost, Bagging, Random Forest and Vote algorithms.

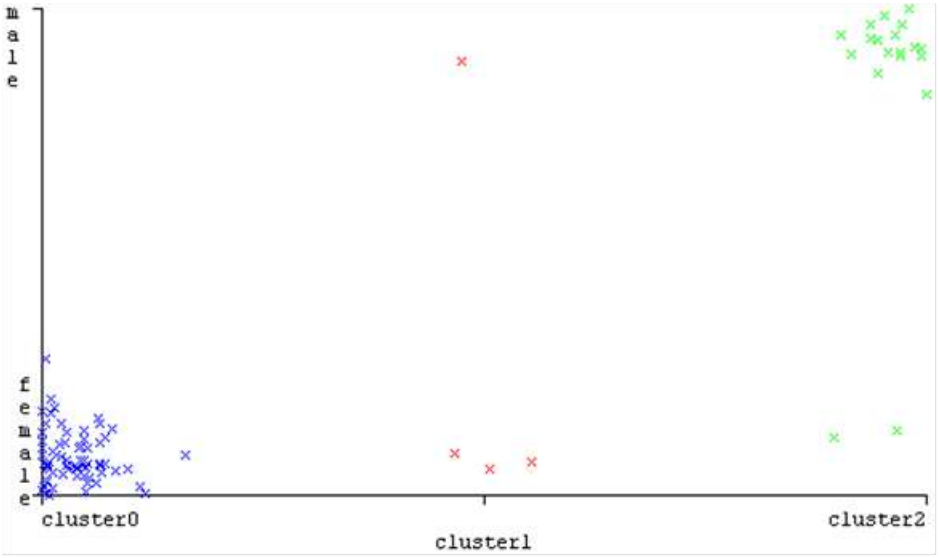

Figure 14: Plot of cluster distribution applying the Expectation Maximization algorithm depending on the sex attribute.

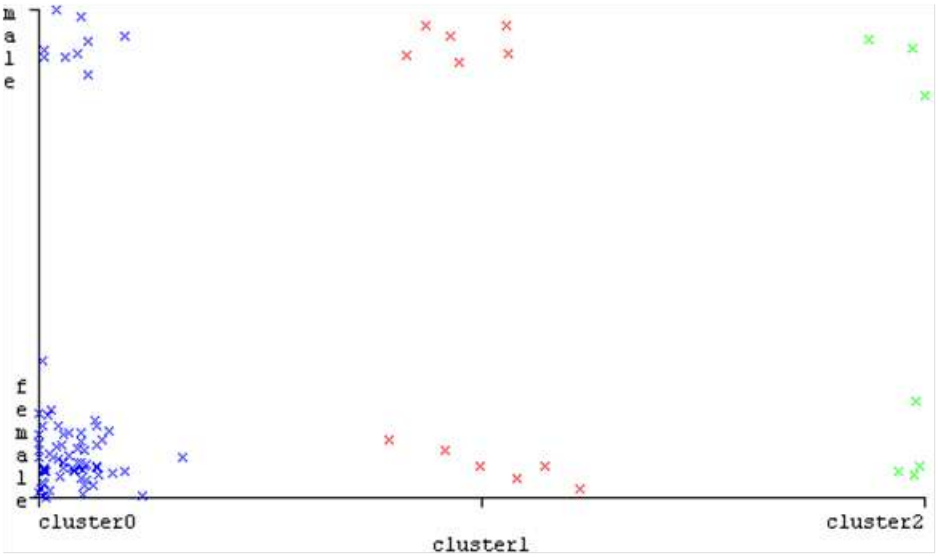

Figure 15: Plot of cluster distribution applying the Farthest First algorithm depending on the sex attribute.

The results are shown in tables 8, 9, 10, 11, 12, 13.

According to the results that are reflected in tables 8-13, the highest per cent of correctly classified instances both by the results of the training model as well as by the prediction results are received while applying the Bagging (classification algorithm classifiers.trees.REPTree) and Random Forest algorithms, with 94.4882% (testing – 96.3303%) and 96.8504% (testing – 99.0826%) respectively. In this case, overfitting is not observed as the stated models demonstrate higher

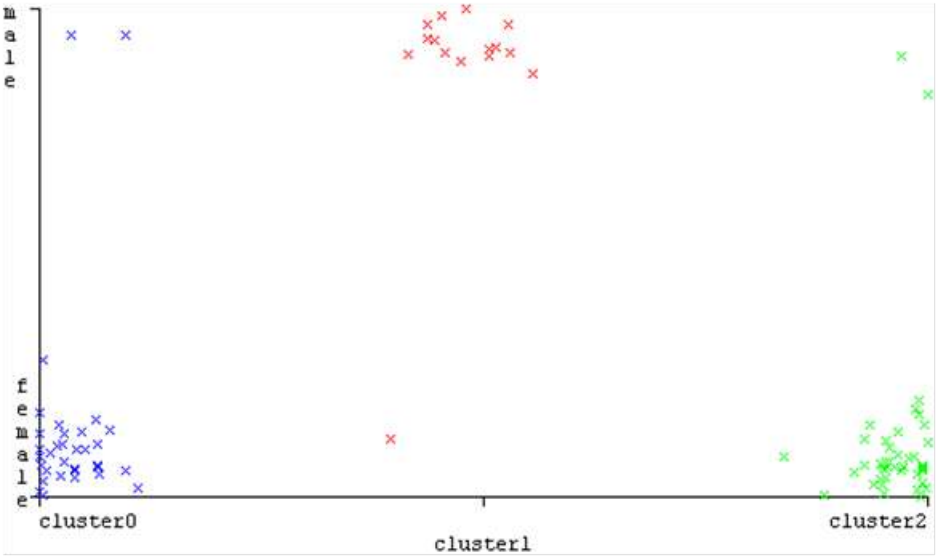

Figure 16: Plot of cluster distribution applying the K-Means algorithm depending on the sex attribute.

Table 8
Evaluation of the results of the work of WEKA ensemble classification training models.

Ensemble classification algorithm scheme	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
weka.classifiers. meta.AdaBoostM1	224 (88.189%)	30 (11.811%)	0.7583	0.1309	0.2355	35.9359%	55.2562%
weka.classifiers. meta.Bagging	240 (94.4882%)	14 (5.5118%)	0.8962	0.0728	0.1765	19.9917%	41.3999%
weka.classifiers. trees.RandomForest	246 (96.8504%)	8 (3.1496%)	0.9411	0.0597	0.1391	16.375%	32.6236%
weka.classifiers. meta.Vote	152 (59.8425%)	102 (40.1575%)	0	0.3643	0.4263	100%	100%

efficiency on test data rather than on training data. In addition, these models demonstrate the highest Kappa statistic and ROC Area indexes. At the same time, the best results are received while using the Random Forest algorithm. The results received in applying the Ada Boost (classifiers.trees.DecisionStump) model are somewhat worse by all the criteria but are still acceptable. According to the indexes provided in tables 8 and 11, the worst results are received in the course of the application of the Vote (classifiers.rules.ZeroR) model.

According to the Mean absolute error (MAE), the data forecast that is close to the actual results both in the process of learning as well as in the process of testing was built using the Random Forest 0.0597 (testing – 0.0405) and Bagging 0.0728 (testing – 0.0478) models; the worst result according to this indicator is received in the course of application of the Vote

Table 9
Detailed Accuracy by Class of the WEKA ensemble classification training models

Ensemble classification algorithm scheme	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC	Area Class
weka.classifiers.meta.AdaBoostM1	0.973	0.000	1.000	0.973	0.986	0.981	0.978	0.982	SR
	1.000	0.294	0.835	1.000	0.910	0.768	0.958	0.961	NR
	0.000	0.000	-	0.000	-	-	0.922	0.620	IR
Weighted Average	0.882	0.176	-	0.882	-	-	0.960	0.929	
weka.classifiers.meta.Bagging	0.973	0.000	1.000	0.973	0.986	0.981	0.975	0.981	SR
	0.980	0.108	0.931	0.980	0.955	0.886	0.972	0.965	NR
	0.679	0.013	0.864	0.679	0.760	0.741	0.975	0.852	IR
Weighted Average	0.945	0.066	0.944	0.945	0.943	0.898	0.974	0.957	
weka.classifiers.trees.Random Forest	0.973	0.000	1.000	0.973	0.986	0.981	0.998	0.996	SR
	0.993	0.069	0.956	0.993	0.974	0.935	0.994	0.995	NR
	0.821	0.004	0.958	0.821	0.885	0.875	0.995	0.971	IR
Weighted Average	0.969	0.042	0.969	0.969	0.968	0.942	0.995	0.993	
weka.classifiers.meta.Vote	0.000	0.000	-	0.000	-	-	0.484	0.285	SR
	1.000	1.000	0.598	1.000	0.749	-	0.475	0.586	NR
	0.000	0.000	-	0.000	-	-	0.466	0.103	IR
Weighted Average	0.598	0.598	-	0.598	-	-	0.477	0.445	

0.3643 (testing – 0.3569) model. Approximately twice the higher MAE value was received while building and testing the Ada Boost model, which is 0.1309 and 0.1029, respectively.

The Root mean squared error (RMSE) values also indicate the supremacy of the Random Forest 0.1391 (testing – 0.0964) and Bagging 0.1765 (testing – 0.1362) algorithms. The worst value was received due to building a model based on Vote 0.4263 (testing – 0.4176).

According to the Relative absolute error (RAE) and Root relative squared error (RRSE), the assessment prioritisation of classification models is preserved with Random Forest and Bagging. It should be noted that the worst indexes are received as a result of classification using the M model (RAE=100%, RSE=100%), which characterises an almost random prediction.

4. Conclusion

In this section, we summarise the key findings derived from our study, which aimed to explore the fields of application and empirically compare ensemble classification and clustering methods within the WEKA machine learning system. Our investigation primarily focused on identifying signs of internet addiction (IA) related disorders among students majoring in Computer Sciences. The following conclusions have been drawn:

1. Through empirical comparisons involving the Expectation Maximization, Farthest First, and K-Means algorithms within the WEKA machine learning system, we successfully

Table 10
Table of confusion matrix of WEKA ensemble classification testing models.

			Actual class		
	Ensemble classification algorithm scheme	Area Class	SR	NR	IR
Predicted class	weka.classifiers.meta.AdaBoostM1	SR	72	2	0
		NR	0	152	0
		IR	0	28	0
	weka.classifiers.meta.Bagging	SR	72	2	0
		NR	0	149	3
		IR	0	9	19
	weka.classifiers.trees.RandomForest	SR	72	2	0
		NR	0	151	1
		IR	0	5	23
	weka.classifiers.meta.Vote	0	74	0	
		NR	0	152	0
		IR	0	28	0

Table 11
Evaluation of the results of testing the WEKA ensemble classification models.

Ensemble classification algorithm scheme	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
weka.classifiers.meta.AdaBoostM1	103 (94.4954%)	6 (5.5046%)	0.8869	0.1029	0.1729	28.8318%	41.399%
weka.classifiers.meta.Bagging	105 (96.3303%)	4 (3.6697%)	0.9276	0.0478	0.1362	13.3815%	32.6122%
weka.classifiers.trees.RandomForest	108 (99.0826%)	1 (0.9174%)	0.982	0.0405	0.0964	11.3566%	23.0819%
weka.classifiers.meta.Vote	65 (59.633%)	44 (40.367%)	0	0.3569	0.4176	100%	100%

developed models for clustering data instances to discern signs of IA disorders among Computer Science students.

2. Implementation of the Expectation Maximization, K-Means, and Farthest First algorithms each formed three distinct clusters. These clusters revealed that a notable characteristic among respondents was the centralisation of the Internet in their psychological reality, often at the expense of other life interests and daily activities. Furthermore, the Expectation Maximization algorithm identified a cluster exhibiting behaviour control disorders associated with online gaming, indicating a higher risk group for IA-related disorders.
3. While the Expectation Maximization, Farthest First, and K-Means algorithms differed in their underlying models, they produced relatively similar clusters regarding characteristic

Table 12
Detailed Accuracy by Class of testing the WEKA ensemble classification models.

Ensemble classification algorithm scheme	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC	Area Class
weka.classifiers.meta.AdaBoostM1	0.974	0.000	1.000	0.974	0.987	0.980	0.983	0.984	SR
	1.000	0.136	0.915	1.000	0.956	0.889	0.976	0.971	NR
	0.000	0.000	-	0.000	-	-	0.946	0.540	IR
Weighted Average	0.945	0.081	-	0.945	-	-	0.977	0.956	
weka.classifiers.meta.Bagging	0.974	0.000	1.000	0.974	0.987	0.980	1.000	1.000	SR
	0.985	0.068	0.955	0.985	0.970	0.924	0.995	0.997	NR
	0.600	0.010	0.750	0.600	0.667	0.657	0.975	0.750	IR
Weighted Average	0.963	0.041	0.962	0.963	0.962	0.932	0.996	0.987	
weka.classifiers.trees.Random Forest	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	SR
	1.000	0.023	0.985	1.000	0.992	0.981	0.999	1.000	NR
	0.800	0.000	1.000	0.800	0.889	0.890	0.998	0.967	IR
Weighted Average	0.991	0.014	0.991	0.991	0.990	0.984	0.999	0.998	
weka.classifiers.meta.Vote	0.000	0.000	-	0.000	-	-	0.500	0.358	SR
	1.000	1.000	0.596	1.000	0.747	-	0.500	0.596	NR
	0.000	0.000	-	0.000	-	-	0.500	0.046	IR
Weighted Average	0.596	0.596	-	0.596	-	-	0.500	0.486	

features, demonstrating optimised clustering. However, the clusters varied in terms of their models, including the number of data instances in each cluster, their structures, and the values of attribute centroids.

4. Assessing the validity of clustering using various indices suggests that the K-Means and Farthest First algorithms may yield less favourable clustering results than the Expectation Maximization algorithm.
5. Respondents were categorised into three groups: Significant Risk (SR), Insignificant Risk (IR), and No Risk (NR). This categorisation allows for a preliminary assessment of IA development risks based on the extent of the Internet's influence on a person's psyche. The SR group reflects deep and maladaptive internet influence, while the IR group shows more superficial and adaptive effects. The NR group indicates the absence of IA risks.
6. The Random Forest algorithm emerged as the most accurate model for classifying respondents, closely followed by the Bagging algorithm (classifiers.trees.REPTree). The Ada Boost algorithm (classifiers.trees.DecisionStump) yielded slightly lower classification scores. Notably, the Vote algorithm (classifiers.rules.ZeroR) needed to prove more suitable, indicating a need for additional modifications.
7. Our data set analysis using ensemble classification and clustering methods suggests their suitability for detecting IA disorders and identifying respondent groups displaying signs of IA-related disorders among Computer Science students.

Table 13
Table of confusion matrix of testing the WEKA ensemble classification models.

		Actual class			
	Ensemble classification algorithm scheme	Area Class	SR	NR	IR
Predicted class	weka.classifiers.meta.AdaBoostM1	SR	38	1	0
		NR	0	65	0
		IR	0	5	0
	weka.classifiers.meta.Bagging	SR	38	1	0
		NR	0	64	1
		IR	0	2	3
	weka.classifiers.trees.RandomForest	SR	39	0	0
		NR	0	65	0
		IR	0	1	4
	weka.classifiers.meta.Vote	0	39	0	
		NR	0	65	0
		IR	0	5	0

8. These findings underscore the potential of applying intelligent data analysis in medical research through machine learning systems. These methods can form the foundation for developing new approaches to processing large medical data sets and making informed decisions in the field.

These results emphasise the importance of unconventional approaches to data analysis in contemporary medicine, where complex methodologies and ensemble techniques can effectively process vast digital data sets. Our findings contribute to the understanding and identification of IA-related disorders among Computer Science students and may inform the development of preventive measures and services to combat IA.

References

[1] Abbott, D.A., Cramer, S.L. and Sherrets, S.D., 1995. Pathological Gambling and the Family: Practice Implications. *Families in Society*, 76(4), pp.213–219. Available from: https://doi.org/10.1177/104438949507600402.

[2] Anokhin, P.K., 1968. Cybernétique, neurophysiologie et psychologie. *Social Science Information*, 7(1), pp.169–197. Available from: https://doi.org/10.1177/053901846800700115.

[3] Antoniuk, D.S., Vakaliuk, T.A., Didkivskyi, V.V., Vizghalov, O., Oliinyk, O.V. and Yanchuk, V.M., 2021. Using a business simulator with elements of machine learning to develop personal finance management skills. In: V. Ermolayev, A.E. Kiv, S.O. Semerikov, V.N. Soloviev and A.M. Striuk, eds. *Proceedings of the 9th Illia O. Teplytskyi Workshop on Computer Simulation in Education (CoSinE 2021) co-located with 17th International Conference on ICT in Education, Research, and Industrial Applications: Integration, Harmonization, and Knowledge Transfer (ICTERI 2021), Kherson, Ukraine, October 1, 2021*. CEUR-WS.org, *CEUR Workshop Proceedings*, vol. 3083, pp.59–70. Available from: https://ceur-ws.org/Vol-3083/paper131.pdf.

[4] Antoniuk, D.S., Vakaliuk, T.A., Didkivskyi, V.V. and Vizghalov, O.Y., 2022. Development of a simulator to determine personal financial strategies using machine learning. *CEUR Workshop Proceedings*, 3077, pp.12–26.
[5] Balatskiy, E.V., 2008. Vitalnyye resursy i kontury soznaniya [Vital resources and circuits of consciousness]. *Vestnik Rossiyskoy akademii nauk*, 78(6), pp.531–537.
[6] Breiman, L., 1996. Bagging predictors. *Machine learning*, 24(2), pp.123–140.
[7] Breiman, L., 2001. Random forests. *Machine learning*, 45(1), pp.5–32.
[8] Brito Da Silva, L.E., Melton, N.M. and Wunsch, D.C., 2020. Incremental Cluster Validity Indices for Online Learning of Hard Partitions: Extensions and Comparative Study. *IEEE Access*, 8, pp.22025–22047. Available from: https://doi.org/10.1109/ACCESS.2020.2969849.
[9] Dasgupta, S. and Long, P.M., 2005. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4), pp.555–569. Available from: https://doi.org/10.1016/j.jcss.2004.10.006.
[10] Derbentsev, V., Matviychuk, A. and Soloviev, V.N., 2020. Forecasting of Cryptocurrency Prices Using Machine Learning. In: L. Pichl, C. Eom, E. Scalas and T. Kaizoji, eds. *Advanced Studies of Financial Technologies and Cryptocurrency Markets*. Singapore: Springer Singapore, pp.211–231. Available from: https://doi.org/10.1007/978-981-15-4498-9_12.
[11] Derhach, M., 2016. Cyber-Addiction of Students Majoring in Computer Science. *Science and Education*, (7), pp.92–98. Available from: https://doi.org/10.24195/2414-4665-2016-7-16.
[12] Di, Z., Gong, X., Shi, J., Ahmed, H.O.A. and Nandi, A.K., 2019. Internet addiction disorder detection of Chinese college students using several personality questionnaire data and support vector machine. *Addictive Behaviors Reports*, 10, p.100200. Available from: https://doi.org/10.1016/j.abrep.2019.100200.
[13] Fadieieva, L.O., 2021. Enhancing adaptive learning with Moodle's machine learning. *Educational Dimension*, 5, p.1–7. Available from: https://doi.org/10.31812/ed.625.
[14] Frankl, V.E., 1985. *Man's search for meaning*. Simon and Schuster.
[15] Freund, Y. and Schapire, R.E., 1996. Experiments with a New Boosting Algorithm. *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., ICML'96, p.148–156.
[16] Hsieh, W.H., Shih, D.H., Shih, P.Y. and Lin, S.B., 2019. An ensemble classifier with case-based reasoning system for identifying internet addiction. *International journal of environmental research and public health*, 16(7), p.1233. Available from: https://doi.org/10.3390/ijerph16071233.
[17] Huizinga, J., 2016. Homo Ludens: A Study of the Play-Element in Culture. *Angelico press*.
[18] Hussain, R.G., Ghazanfar, M.A., Azam, M.A., Naeem, U. and Ur Rehman, S., 2019. A performance comparison of machine learning classification approaches for robust activity of daily living recognition. *Artificial Intelligence Review*, 52(1), pp.357–379. Available from: https://doi.org/10.1007/s10462-018-9623-5.
[19] Ji, H.M., Chen, L.Y. and Hsiao, T.C., 2019. Real-time detection of internet addiction using reinforcement learning system. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. pp.1280–1288. Available from: https://doi.org/10.1145/3319619.3326882.
[20] Keng, B., 2016. The Expectation-Maximization Algorithm. Available from: http://bjlkeng.github.io/posts/the-expectation-maximization-algorithm.

[21] Kittler, J., Hatef, M., Duin, R.P. and Matas, J., 1998. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3), pp.226–239.

[22] Kiv, A., Hryhoruk, P., Khvostina, I., Solovieva, V., Soloviev, V.N. and Semerikov, S., 2020. Machine learning of emerging markets in pandemic times. In: A. Kiv, ed. *Proceedings of the Selected Papers of the Special Edition of International Conference on Monitoring, Modeling & Management of Emergent Economy (M3E2-MLPEED 2020), Odessa, Ukraine, July 13-18, 2020*. CEUR-WS.org, *CEUR Workshop Proceedings*, vol. 2713, pp.1–20. Available from: https://ceur-ws.org/Vol-2713/paper00.pdf.

[23] Kiv, A., Semerikov, S., Soloviev, V.N., Kibalnyk, L., Danylchuk, H. and Matviychuk, A., 2019. Experimental Economics and Machine Learning for Prediction of Emergent Economy Dynamics. In: A. Kiv, S. Semerikov, V.N. Soloviev, L. Kibalnyk, H. Danylchuk and A. Matviychuk, eds. *Proceedings of the Selected Papers of the 8th International Conference on Monitoring, Modeling & Management of Emergent Economy, M3E2-EEMLPEED 2019, Odessa, Ukraine, May 22-24, 2019*. CEUR-WS.org, *CEUR Workshop Proceedings*, vol. 2422, pp.1–4. Available from: https://ceur-ws.org/Vol-2422/paper00.pdf.

[24] Kiv, A.E., Soloviev, V.N., Semerikov, S.O., Danylchuk, H.B., Kibalnyk, L.O., Matviychuk, A.V. and Striuk, A.M., 2021. Machine learning for prediction of emergent economy dynamics. In: A.E. Kiv, V.N. Soloviev and S.O. Semerikov, eds. *Proceedings of the Selected and Revised Papers of 9th International Conference on Monitoring, Modeling & Management of Emergent Economy (M3E2-MLPEED 2021), Odessa, Ukraine, May 26-28, 2021*. CEUR-WS.org, *CEUR Workshop Proceedings*, vol. 3048, pp.i–xxxi. Available from: https://ceur-ws.org/Vol-3048/paper00.pdf.

[25] Klochko, O. and Fedorets, V., 2019. An empirical comparison of machine learning clustering methods in the study of Internet addiction among students majoring in Computer Sciences. *CEUR Workshop Proceedings*, 2546, pp.58–75.

[26] Klochko, O., Fedorets, V., Tkachenko, S. and Maliar, O., 2020. The Use of Digital Technologies for Flipped Learning Implementation. In: O. Sokolov, G. Zholtkevych, V. Yakovyna, Y. Tarasich, V. Kharchenko, V. Kobets, O. Burov, S. Semerikov and H. Kravtsov, eds. *Proceedings of the 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, Kharkiv, Ukraine, October 06-10, 2020*. CEUR-WS.org, *CEUR Workshop Proceedings*, vol. 2732, pp.1233–1248. Available from: https://ceur-ws.org/Vol-2732/20201233.pdf.

[27] Klochko, O.V., 2019. *Matematychne modeliuvannia system i protsesiv v osviti/pedahohitsi [Mathematical modeling of systems and processes in education/pedagogy]*. Druk.

[28] Krämer, J., Schreyögg, J. and Busse, R., 2019. Classification of hospital admissions into emergency and elective care: a machine learning approach. *Health Care Management Science*, 22(1), pp.85–105. Available from: https://doi.org/10.1007/s10729-017-9423-5.

[29] Kuncheva, L.I., 2014. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

[30] Lee, T.H., Ullah, A. and Wang, R., 2020. Bootstrap Aggregating and Random Forest. In: P. Fuleky, ed. *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*. Cham: Springer International Publishing, pp.389–429. Available from: https://doi.org/10.1007/978-3-030-31150-6_13.

[31] Leontyev, D.A., 2017. *Psikhologiya smysla: priroda. stroyeniye i dinamika smyslovoy realnosti*

[The psychology of meaning: nature, structure and dynamics of meaningful reality]. Litres.

[32] Linoff, G.S. and Berry, M.J.A., 2011. *Data mining techniques: for marketing, sales, and customer relationship management.* John Wiley & Sons.

[33] Moshtaghi, M., Bezdek, J.C., Erfani, S.M., Leckie, C. and Bailey, J., 2019. Online cluster validity indices for performance monitoring of streaming data clustering. *International Journal of Intelligent Systems*, 34(4), pp.541–563. Available from: https://doi.org/10.1002/int.22064.

[34] Pacol, C.A. and Palaoag, T.D., 2020. Enhancing Sentiment Analysis of Textual Feedback in the Student-Faculty Evaluation using Machine Learning Techniques. *3rd International Conference On Academic Research in Science, Technology and Engineering*. Available from: https://www.dpublication.com/wp-content/uploads/2020/11/3-3003.pdf.

[35] Santos, S.G.T.d.C. and Barros, R.S.M. de, 2020. Online AdaBoost-based methods for multiclass problems. *Artificial Intelligence Review*, 53(2), pp.1293–1322. Available from: https://doi.org/10.1007/s10462-019-09696-6.

[36] Semerikov, S.O., Vakaliuk, T.A., Mintii, I.S., Hamaniuk, V.A., Soloviev, V.N., Bondarenko, O.V., Nechypurenko, P.P., Shokaliuk, S.V., Moiseienko, N.V. and Ruban, V.R., 2021. Development of the computer vision system based on machine learning for educational purposes. *Educational Dimension*, 5, p.8–60. Available from: https://doi.org/10.31812/educdim.4717.

[37] Souri, A., Ghafour, M.Y., Ahmed, A.M., Safara, F., Yamini, A. and Hoseyninezhad, M., 2020. A new machine learning-based healthcare monitoring model for student's condition diagnosis in Internet of Things environment. *Soft Computing*, 24(22), pp.17111–17121. Available from: https://doi.org/10.1007/s00500-020-05003-6.

[38] Subasi, A., Kevric, J. and Abdullah Canbaz, M., 2019. Epileptic seizure detection using hybrid machine learning methods. *Neural Computing and Applications*, 31(1), pp.317–325. Available from: https://doi.org/10.1007/s00521-017-3003-y.

[39] Sudakov, K.V., 2011. *Funktsionalnyye sistemy [Functional systems)].* Izdatelstvo RAMN.

[40] Suma, S.N., Nataraja, P. and Sharma, M.K., 2021. Internet Addiction Predictor: Applying Machine Learning in Psychology. *Advances in Artificial Intelligence and Data Engineering*. Springer, pp.471–481.

[41] Tuysuzoglu, G. and Birant, D., 2020. Enhanced bagging (eBagging): A novel approach for ensemble learning. *The International Arab Journal of Information Technology*, 17(4), pp.515–528. Available from: https://doi.org/10.34028/iajit/17/4/10.

[42] Wallis, D., 1997. Just click no: Talk story about Dr. Ivan K. Goldberg and the internet addiction disorder. *The New Yorker*. Available from: http://www.newyorker.com/magazine/1997/01/13/just-click-no.

[43] Weka 3: Machine Learning Software in Java, 2023. Available from: http://old-www.cms.waikato.ac.nz/~ml/weka/.

[44] Young, K.S., 1998. *Caught in the net: How to recognize the signs of internet addiction–and a winning strategy for recovery.* John Wiley & Sons.

[45] Young, K.S., 1998. Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior*, 1(3), pp.237–244. Available from: https://doi.org/10.1089/cpb.1998.1.237.

[46] Yuryeva, L.N. and Bolbot, T.Y., 2006. *Kompyuternaya zavisimost: formirovaniye, diagnostika, korrektsiya i profilaktika [Computer addiction: formation, diagnostics, correction and*

prevention]. Dnepropetrovsk: Porogi. Available from: http://kingmed.info/media/book/3/2673.pdf.

[47] Zahorodko, P.V., Modlo, Y.O., Kalinichenko, O.O., Selivanova, T.V. and Semerikov, S.O., 2020. Quantum enhanced machine learning: An overview. *CEUR Workshop Proceedings*, 2832, pp.94–103.

[48] Zahorodko, P.V., Semerikov, S.O., Soloviev, V.N., Striuk, A.M., Striuk, M.I. and Shalatska, H.M., 2021. Comparisons of performance between quantum-enhanced and classical machine learning algorithms on the IBM Quantum Experience. *Journal of Physics: Conference Series*, 1840(1), p.012021. Available from: https://doi.org/10.1088/1742-6596/1840/1/012021.

[49] Zelinska, S., 2020. Machine learning: Technologies and potential application at mining companies. *E3S Web of Conferences*, 166, p.03007. Available from: https://doi.org/10.1051/e3sconf/202016603007.

[50] Zhong, Y., Yang, H., Zhang, Y. and Li, P., 2020. Online random forests regression with memories. *Knowledge-Based Systems*, 201, p.106058.