



# The Use of Ensemble Classification and Clustering Methods of Machine Learning in the Study of Internet Addiction of Students

Oksana V. Klochko<sup>1</sup><sup>a</sup>, Vasyl M. Fedorets<sup>2</sup><sup>b</sup>, Vitalii I. Klochko<sup>3</sup><sup>c</sup> and Maryna V. Kormer<sup>4</sup><sup>d</sup>

<sup>1</sup>Vinnitsia Mykhailo Kotsiubynskyi State Pedagogical University, 32 Ostrozhskego Str., Vinnitsia, 21100, Ukraine

<sup>2</sup>Vinnitsia Academy of Continuing Education, 13 Hrushevskoho Str., Vinnitsia, 21050, Ukraine

<sup>3</sup>Vinnitsia National Technical University, 95 Khmelnytsky Highway, Vinnitsia, 21021, Ukraine

<sup>4</sup>State University of Economics and Technology, 5 Stepana Tilhy Str., Kryvyi Rih, 50006, Ukraine

**Keywords:** Machine Learning, Clustering Classification, Internet Addiction, Detection of Internet Addiction, Internet Disorders, Internet Addiction of Students, Expectation Maximization, Farthest First, K-Means, AdaBoost, Bagging, Random Forest, Vote.


**Abstract:** One of the relevant current vectors of study in machine learning is the analysis of the application peculiarities for methods of solving a specific problem. We will study this issue on the example of methods of solving the clustering and classification problem. Currently, we have a considerable number of machine learning algorithms – e.g. Expectation Maximization, Farthest First, K-Means, Expectation-Maximization, Hierarchical Clustering, Support vector machines, K-nearest neighbor, Logistic regression, Random Forest etc. – which can be used for clustering and classification. However, not all methods can be used for solving a specific task. The article describes the technology of empirical comparison of methods of clustering and classification problems solving using WEKA free software for machine learning. Empirical comparison of data clustering methods was based on the results of a survey conducted among students majoring in Computer Studies and dedicated to detecting signs of Internet Addiction (IA) (Internet Addiction is a behavioural disorder that occurs due to Internet misuse). As a continuation of the study of Internet Addiction of students, a survey of students of other specialties was conducted. Ensemble methods of machine learning classification were used to analyze these data. Empirical comparison of clustering algorithms (Expectation Maximization, Farthest First and K-Means) and ensemble classification algorithms (AdaBoost, Bagging, Random Forest and Vote) with the application of the WEKA machine learning system had the following results: it described the peculiarities of application of these methods in feature clustering and classification, the authors developed data instances' clustering and classification models to detect signs of Internet addiction among students, the study concludes that these methods may be applicable to development of models detecting respondents with signs of IA related disorders and risk groups.


## 1 INTRODUCTION


One of the areas of research, particularly, such as education, healthcare and life safety, is the empirical analysis of methods of solving a specific problem with the using of the machine learning (Zahorodko et al., 2021; Zelinska, 2020). Let us study this issue on the example of methods of solving the clustering and classification problems (Tarasenko et al., 2019).


Clustering methods are statistic methods of data analysis that enable people to group the given selection of data samples into clusters, classes, taxons depending on the value of their attributes; each of these groups has certain characteristics. The main idea is to use several clustering methods in order to carry out an empirical comparison study and determine which methods ensure the most optimal data grouping while solving a specific problem.

Machine learning classifies clustering problems as problems for unsupervised learning. Currently, there is a considerable number of machine learning algorithms that can be used for clustering, for instance, Expectation Maximization, Farthest First, K-Means,

<sup>a</sup> <https://orcid.org/0000-0002-6505-9455>

<sup>b</sup> <https://orcid.org/0000-0001-9936-3458>

<sup>c</sup> <https://orcid.org/0000-0002-9415-4451>

<sup>d</sup> <https://orcid.org/0000-0002-6509-0794>

K-Medians, Hierarchical Clustering etc. But not all of them are suitable for solving a specific problem. Data clustering algorithms differ by the cluster model type, the algorithm model type, the nesting hierarchy of clusters, the way of implementation depending on the data set etc. Because of this, there are also certain requirements to the data set parameters.

Classification tasks are considered to be a different group of machine learning tasks – namely, supervised learning. However, nowadays, there are developed classification algorithms based on a combination of supervised learning and unsupervised learning (e.g., Learning Vector Quantization) the classification algorithm in machine learning is built on the basis of the preset finite number of objects divided into groups and allows to classify an arbitrary object, if it is unknown which group it belongs to. There is a considerable number of machine learning algorithms that can be applied to solve classification tasks, e.g. Support vector machines (SVMs), Logistic regression, K-nearest neighbor, Linear discriminant analysis (LDA), etc. However, currently, one can get more accurate results by using ensemble methods.

Popular software products used in machine learning include TensorFlow, WEKA, MATLAB, MXNet, Torch, PyTorch, Microsoft Azure Machine Learning Studio and others.

In this study, we use the WEKA (Waikato Environment for Knowledge Analysis) free machine learning software (Weka, 2021). The free WEKA machine learning system gives direct access to the library of implemented algorithms written in Java.

Analysis of contemporary studies and publications shows that the issue of analysis and selection of the machine learning method, which would be optimal for processing a concrete data set, is popular in the scientific circles. A considerable number of these studies is dedicated to the application of machine learning methods in the fields of education, healthcare and life safety.

In healthcare and education sentiment analysis becomes more and more popular. Thus, Pacol and Palaoag (Pacol and Palaoag, 2020) conducted the sentiment analysis of the Textual Feedback of students regarding the work of the professors using Machine Learning Techniques. In the sentiment classification study, the Random Forest algorithm proved to be most effective; it proved to be more effective than base models of Support vector machines, Naive Bayes, Logistic regression algorithms and their ensembles (Pacol and Palaoag, 2020).

Klochko et al. (Klochko et al., 2020) applied clustering algorithms of machine learning for the analysis of typical mistakes pupils make, the selection

and adaptation of the content of learning to concrete groups of pupils that was then supposed to be used for flipped learning with the use of virtual learning environment. The analysis was done through the comparison of clusters, defined by learning results demonstrated by the pupils, using Canopy, Expectation Maximization and Farthest First algorithms.

Souri et al. (Souri et al., 2020) suggested a model based on the Internet of Things technologies for monitoring the indicators of students' health in order to detect biological and behavioral changes. The developed model, when used together with the Support Vector Machine (SVM) algorithm, reaches the highest accuracy of 99.1% (Souri et al., 2020).

Hussain et al. (Hussain et al., 2019) studied the application of the machine learning classification methods to find ways to ensure independent daily living of people who have Alzheimer's disease. The idea of the study is to analyze the data registered by different equipment in order to determine the changes in a person's behavior that are relevant for the daily life and social interaction. The paper gives a comparison of the efficiency levels of five machine learning classification techniques used for the recognition of a person's activity (and his/her psychological status). Experimental findings show that compared to traditional methodologies, these approaches give better results in determining the activity of the person and his/her psychological and behavioral peculiarities.

Krämer et al. (Krämer et al., 2019) studied the speed and efficiency of medical aid provision using the databases of the Hospital ER. Applying the Random forest algorithm, the authors developed the model based on the data about the patient's provisional diagnosis. The use of the controlled machine learning method and model training based on the opinion of a specialized doctor allowed them to achieve high forecasting accuracy (96%) as well as the area under the receiver operating curve ( $>0.99$ ).

Subasi et al. (Subasi et al., 2019) developed a hybrid model of detecting epileptic fits using the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to determine the optimal parameters of application of the Support Vector Machine (SVM) algorithm. The hybrid algorithm that they suggested can demonstrate data set classification accuracy of up to 99.38%.

A considerable number of papers appeared, which are dedicated to diagnosing Internet addiction (IA) and studying the mechanisms of this disorder among various social groups. The appearance and use of the Internet has many benefits. However, at the same time, disorders related to pathological use of the Internet are becoming a social as well as a psychological

problem. Currently, we face an important psychological, sociocultural and educational issue of detection and prevention of certain pathologies and steady pre-morbid conditions (state before the disease) caused by inadequate Internet use. Cases of IA were first mentioned in 1995 and attracted considerable attention. Issues related to this one became the research subject of Yuryeva and Bolbot (Yuryeva and Bolbot, 2006), Derhach (Derhach, 2016) and others. Internet Addiction Disorder (IAD) is also called Pathological Internet Use (PIU). The term “Internet Addiction” was first suggested by Ivan K. Goldberg in 1995. He describes net addiction as a specific pathology characterized by a wide spectrum of behavioral and impulse control disorders (lack of control, absence of voluntary regulation) (Abbott et al., 1995). In 1996 Goldberg made the first attempt to determine groups of behavioural and psychological signs and symptoms of IA (Wallis, 1997), namely: tolerance; abstinence syndrome; difficulties in voluntary regulation of Internet-behaviour; increase of time and financial investments in things related to Internet or computer use; a shift of a person’s interests towards Internet-related activities; extensive Internet use that leads to maladjustment. In 1998 Young (Young, 1998a,b) defined IAD as an impulsive-compulsive disorder, which has specific signs or addictions: cyber-sexual addiction, cyber-relationship addiction, net compulsions, information overload and computer addiction. IAD is not officially included into ICD-11 for Mortality and Morbidity Statistics (Version: 09/2020), however, in section 6C51 Gaming disorder the “Gaming disorder” is described as a “pattern of persistent or recurrent gaming behaviour (‘digital gaming’ or ‘video-gaming’), which may be online (i.e., over the Internet)” (ICD-11 for Mortality and Morbidity Statistics, Version: 09/2020).

Even though the problem of IA is becoming more and more relevant, there are not enough scientific papers dedicated to the study of this issue with the help of machine learning methods. Let us look at some of them. On the basis of the Support Vector Machine algorithm, including the C-SVM and v-SVM, and applying the Student’s t-test to the data set of the survey conducted among 2,397 Chinese students, Di et al. (Di et al., 2019) proved the utility of using machine learning methods for detecting and forecasting the risk of IA. Hsieh et al. (Hsieh et al., 2019) suggested using the EMBAR protected system of web-services based on the ensemble classification methods and case-based reasoning to study the IA of the users and prevent the development of this disorder at the initial stages. Ji et al. (Ji et al., 2019) are currently continuing their research, which aims to cre-

ate an IA detector that would work in a real-time mode. The authors suggest studying this issue using an adapted system of continuous real-coded variables (XCSR), which determines the level of Internet addiction (high-risk and low-risk) on the basis of the information about the Internet users using the Chen Internet addiction scale (CIAS) or respiratory instantaneous frequency (IF). Suma et al. (Suma et al., 2021) studied the possibilities for predicting IA based on a set of predictor variables using the Random forest algorithm.

Thus, based on the above presented statement of the problem as well as taking into consideration the insufficient amount of research on the application of machine learning methods to IA diagnosing, we determine the aim of our research, which is to determine the fields of use and conduct an empirical comparison of ensemble classification and clustering methods of machine learning in the study of IA disorder of students.

## 2 SELECTION OF METHODS AND DIAGNOSTICS

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

At 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. 262 students majoring in Computer Sciences 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) (Klochko and Fedorets, 2019). The data set contains the fields described in table 1.

Cluster analysis is one of the tasks of database mining. Cluster analysis is a set of methods of multidimensional observations or objects classification, based on defining the concept of distance between the objects and their subsequent grouping (into clusters, taxons, classes). The selection of a concrete cluster analysis method depends on the purpose of classification (Klochko, 2019). At the same time, one does not need a priori information about the statistical popula-

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 from food, sleep, going to classes, if a 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've spent not enough time playing computer games over the Internet, I constantly wish to play longer	Nominal	yes 178 undefined 61 no 23

@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.

tion. This approach is based on the following presuppositions: objects that have 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 (Klochko, 2019).

Let us look at one of cluster analysis algorithms (Klochko, 2019).

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 standardized Z values with elements:

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

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

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

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

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  $l = 1, 2, \dots, n$ , 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 way of distance measuring gives the same result as the usual Euclidean distance, but in this case the influence of non-systemic large 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 in order to determine the differences existing 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 covariance matrix; this distance measurement gives good results when applied to a concrete data group, but it does not work very 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 independence of random variables, which indicates the distance in the orthogonal space.

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

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

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

As the R matrix is symmetric, i.e.  $\rho_{iv} = \rho_{vi}$ , we may confine ourselves to off-diagonal matrix elements. Using the distance matrix, we can implement the agglomerative hierarchic procedure of cluster analysis. Distances between clusters are determined as the closest or the farthest ones. In the first case, the distance between the clusters is the one between the closest elements of these clusters, in the second case, it is the one between the two farthestmost located. The principle of the work of agglomerative hierarchic procedures lies in a consequent grouping of elements, starting from the ones closest to each other and those that are farther and farther apart. During the first step of the algorithm, every observation  $z_i$  ( $i = 1, 2, \dots, m$ ) is viewed as a separate cluster. Then, during every next step of the work of the algorithm, two closest located clusters are grouped together 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 indicated its belongingness to each cluster. EM methods (Keng, 2016): Maximum Likelihood Estimation (MLE) or Maximum a Posteriori (MAP). Description of the algorithm is shown in figure 2 (Keng, 2016): 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.

K-Means algorithm:

Aims to partition  $n$  observations into  $k$  clusters in such a way 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 minimizing the sum of squares of the distances (Linoff and Berry, 2011) (figure 3).

Farthest First algorithm:

This is a modification of a K-Means algorithm, in which the initial selection of centroids is 2 and higher. Centroids are determined following the remoteness principle, i.e. the point farthest from the rest is selected first. The Farthest First algorithm is described in figure 4 (Dasgupta and Long, 2005).

During the second stage of the survey on the situation with IA among students of various specialties was conducted with the help of Google Forms. 363



0. **Initialization:** Get an initial estimate for parameters  $\theta^0$  (e.g. all the  $\mu_k$ ,  $\sigma_k^2$ , and  $\pi$  variables). In many cases, this can just be a random initialization.
1. **Expectation Step:** Assume the parameters ( $\theta^{t-1}$ ) 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  $\theta^t$  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 (Keng, 2016).

**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 (Linoff and Berry, 2011).

students from different regions of Ukraine took part in the survey. The data set is presented in the ARFF format and 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 objects classification, based on defining the concept of distance between the objects and their subsequent grouping (into clusters, taxons, classes). The selection of a concrete cluster analysis method depends on the purpose of classification (Klochko, 2019). 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 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 (Klochko, 2019).

In order to analyze the IA phenomenon, we divide the respondents into three groups (Significant Risk (SR), Insignificant Risk (IR), 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 on the basis of

detecting and analysis of those IA features, which signify qualitative changes of the psychological status of a personality. The selection of such features is carries out on the basis of traditional understanding of the fact that in-depth psychic changes related to the formation of addictive behavior concern, primarily, vital (Balatskiy, 2008) 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 (Frankl, 1985), vital resources (Balatskiy, 2008) and intentions as well as attitudes and behavioral stereotypes aimed at survival and life preservation. They are linked with the vital “foundations” of life itself.

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<sup>st</sup> SR) and “When I cannot use the Internet, I feel anxiety, irritation” (2<sup>nd</sup> SR). The 1<sup>st</sup> SR seems to be more important as food and sleep are system organizing 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 (Frankl, 1985;

```

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

Pick a point and label it 1.

For  $i = 2, 3, \dots, n$ 
    Find the point furthest from  $\{1, 2, \dots, i - 1\}$  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 (Dasgupta and Long, 2005). Take the distance from a point  $x$  to a set  $S$  to be  $d(x, S) = \min_{y \in S} d(x, y)$  (Dasgupta and Long, 2005).

```

@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.

Leontyev, 2017), but also the deformation of very vital “foundation” of a personality. The stated issues of food and sleep indirectly reflect the existential problems of a person. This is caused by the fact 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<sup>st</sup> SR) question, the “can give up classes” part is an important social and personality oriented aspect. If the answer is positive, that means that the content embedded in the afore mentioned fragment is ignored and desensitized. This discloses the presence of desensitization and depreciation of the possible socio-economically “settled” future and a conscious self-limitation in the field of self-actualization in studying and professional activity. Another relevant point is ignoring communica-

tion, 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 desensitized. The second (2<sup>nd</sup> 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 occurs as well as the corresponding changes in the value-conceptual field. This is what V. Frankl (Frankl, 1985) 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 particularized through actualization in the 2<sup>nd</sup> SR question of the irritation phenomenon (“... feel ... irritation”).

In general, the 2<sup>nd</sup> SR question supplements, particularizes and “strengthens” the deficit and “narrowing” of vitality, life creativity, nature corresponding existence, healthy life preservation instinct (food, sleep, communication). The stated vital resources and life creativity, which are actualized and problematized in the 1<sup>st</sup> SR and 2<sup>nd</sup> SR question, act as a complex diagnostic sub-system. It is aimed at detecting systemic, comprehensive stable, in-depth, vital, personal-psychological problems (i.e. disorders). The stated problems may develop and together transform into AI.

The 1<sup>st</sup> SR and 2<sup>nd</sup> 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<sup>st</sup> SR and 2<sup>nd</sup> SR). The questions, which represent the SR group indicate that the Internet has “penetrated” deep into the consciousness and into the core of a personality, in his/her vitality, into human existence (figure 6).

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 from food, sleep, 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've spent not 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

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 (pre-morbid) state or even the presence of the actual IA pathology whereas 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 the in-depth personality-psychological, vital and existential aspect of IA. 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 characterized (diagnosed) by the presence of three questions out of questions 3 (1<sup>st</sup> IR), 5 (2<sup>nd</sup> IR), 7 (3<sup>rd</sup> IR) and 8 (4<sup>th</sup> IR) of table 2. The 1<sup>st</sup> IR question discloses 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 considerably high level or significance and even value of the Internet in the life of a person. The 2<sup>nd</sup> IR question characterizes the peculiarity of the contemporary Internet-culture. It reflects the fact that the person has an actualized orientation-searching reflex and a corresponding search and cognitive behavior, and not just that possibility of IA development. While disclosing the peculiarities of modern-day Internet communication, the 3<sup>rd</sup> IR question also partially characterizes the problem of the insufficiently developed communicative competence and communicative culture of a personality, which, in turn, is effectively compensated with Internet-communication.

As for the 4<sup>th</sup> 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



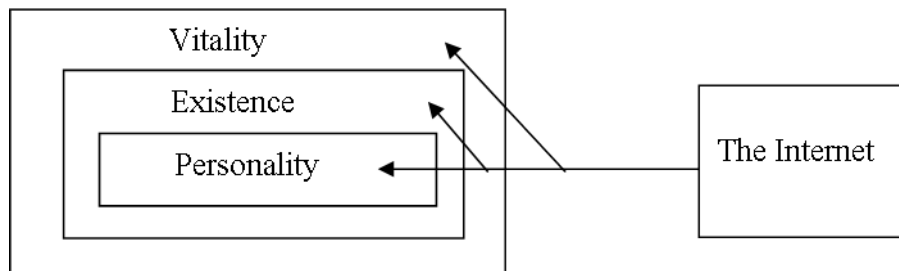


Figure 6: “Penetration” of the Internet into the core of a personality, into vitality, into existence, which illustrates the psychological mechanisms of risk formation in the SR group.

the presence of insufficient risks. At the same time, in its essence, a person is a creature that plays – Homo ludens (Huizinga, 2016). What is important, is the fact that at an early childhood a game is a specific integrative and integrating form of activity and the essence of human existence. The states gaming essence of a person can manifest itself in course of solving complicated tasks, studying as well as during leisure time. A computer game has a considerable mobilizing, emotional, orientation-search potential. A computer game can become addictive, both 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-up develop computer game addiction as a consequence of the fact that they did not have a chance to fully realize 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 (Anokhin, 1968) study on functional systems (Sudakov, 2011). At the same time, constant interest in playing computer games poses a certain threat of the development of a computer game 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<sup>st</sup> IR, 2<sup>nd</sup> IR, 3<sup>rd</sup> IR and 4<sup>th</sup> IR questions, this indicates a certain risk of IA development. This is caused by the fact that each question characterizes the influence of the Internet on a certain aspect of a personality: the 1<sup>st</sup> IR – on the need and value-conceptual aspect, the 2<sup>nd</sup> – on the cognitive aspect, including the orientation ability; the 3<sup>rd</sup> – on the communicative aspect, the 4<sup>th</sup> – on the activity component. Thus, actualization of the Internet as a need, value-conceptual, cognitive, communicative and activity phenomenon that is significant for a personal-

ity speaks of its certain “spread” and “rootedness” in the stated aspects (spheres) as well as about its corresponding significance and value. A certain “Internet locus” is formed in various spheres of a personality. Thus, as a result of systemic interiorization 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 behavioral 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 the number of opportunities for forming the summing up and synergy effects, which lead to qualitative changes.

NR is a group of questions, which characterize separate features of Internet influence. The personality demonstrates local, superficial, reactive reactions. That is why the people that fall into this group, the risks are almost absent or are rather insignificant.

The classification ensemble methods of machine learning were used to study the presence of 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 approach of meta-learning in order to best combine a few machine learning models. On the basis of the basic-level models, the algorithm is taught. Using these results, the meta-model learns to better combine the predictions of basic models. Bagging uses multiple teaching of an ensemble of classifiers on random data sets, which is conducted simultaneously but independently from one another. Then, a determined averaging of results is conducted. The results are averaged on the basis of

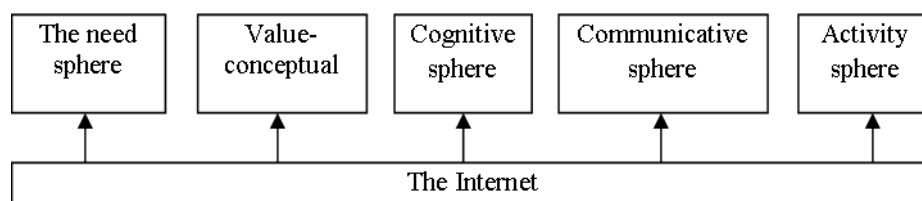


Figure 7: Influence of the Internet on various spheres of human psychich.

a determined strategy. Boosting carries out a consecutive adaptive algorithm teaching. The next algorithm learns through focusing on classification mistakes of the first algorithm. In this research, the authors used such ensemble classification algorithms as AdaBoost, Bagging, Random Forest and Vote.

The WEKA machine learning system uses the Adaboost algorithm.M1 (figure 8) (Freund and Schapire, 1996; Santos and de Barros, 2020). 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 instance that were 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.

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 (Breiman, 1996; Tuysuzoglu and Birant, 2020) (figure 9).

According to Leo Breiman's definition, "a Random Forest is a classifier consisting of a collection of tree-structured classifiers  $\{h(x, \Theta_k), k = 1, \dots\}$  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) (Breiman, 2001; Zhong et al., 2020).

The voting classifier combines different classifiers that learn and are assessed simultaneously (Kittler et al., 1998; Kuncheva, 2014). 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 (Kittler et al., 1998; Kuncheva, 2014). 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 (Kittler et al., 1998; Kuncheva, 2014).

### 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 (Weka, 2021).

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 by the fact these algorithms have long been used to cluster different types of data in many fields and are considered to be effective.

Dunn, DB, SD, CDbw and S\_Dbw were selected as validity indices for testing (da Silva et al., 2019; Moshtaghi et al., 2019) (table 3). In the CDbw index the distance from the point to multitude set in the course of selecting cluster element can be calculated in different ways. In this study, we use the sum of distances of already existing "representatives" of the cluster 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.

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

To cluster the data, we select training/testing using the percentage split option. As a data set for training (model building) we select 66% of data from the set.

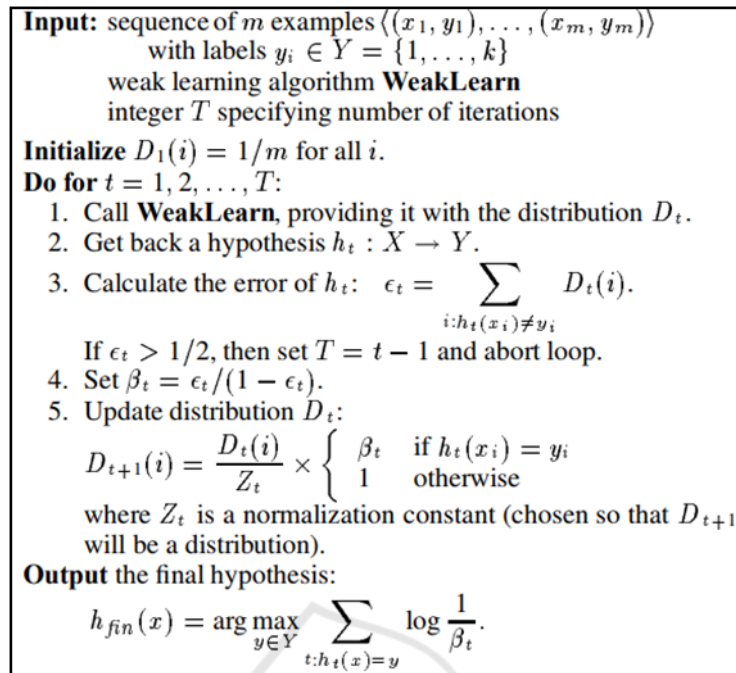


Figure 8: The algorithm AdaBoost.M1 (Freund and Schapire, 1996).

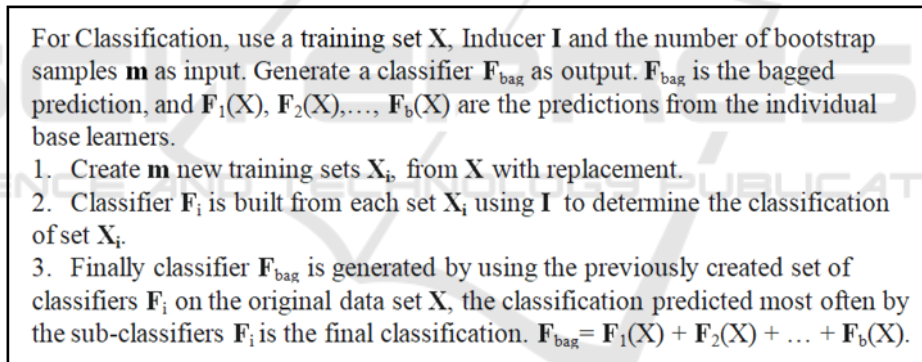


Figure 9: The algorithm Bagging (Breiman, 1996; Lee et al., 2020).

As a data set for testing we select 34% of data from the set. In addition, we select number of clusters “3” in algorithm settings.

We received the following results:

In the course of application of 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 the representatives of this group is that they are unable to 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 behavior, other life interests as well as significance of everyday activities lose their importance. The stated tendencies are linked to IA.

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

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 (Breiman, 2001).

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

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

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 behavior; other life interests as well as 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 the representatives of this group is middle-aged in comparison with

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, predominantly respondents believe they have this addiction. Judging by the values of attributes 4, 5, 6 and 7, the vast majority 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, which was 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; behavioral impulse control disorders related to online gaming. These people are in the risk group for developing IA related disorders.

In the course of application of the Farthest First algorithm, according to the built clustering model based on the training data set, there have also been three clusters 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 mostly female data instances. The representatives of this group cannot imagine their life without the Internet, i.e. there is obvious Internet centration in the psychic reality of a personality. Respondents cannot clearly determine whether they feel either anxiety or irritation if they do not have the possibility to 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, prevalence of Internet relations over actual real interactions, feeling 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 inclination to use the Internet without a concrete purpose. To give an overall characteristic, this group has signs of IA, i.e. behavior 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 do not have the possibility to use the Internet. They are characterized by their undecidedness regarding the vital need to use the Internet; giving up other life interests and everyday activities for the sake of free Internet use; 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 behavior control disorders, related to Internet use. Compared to other groups, they are in the risk group for developing IA related disorders.

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

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 cannot clearly determine whether they have a vital need to use the Internet. As for other indices, respondents state absence of signs of IA related disorders.

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 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, feeling the lack of time spent playing online computer games as well as the systemic need to play longer. The overall characteristic of this cluster is the presence of signs of IA, i.e. behavior control issues related to Internet use, namely, gaming Internet addiction. If compared with other cluster, 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 the course of application of the three algorithms – the Expectation Maximization, Farthest First and K-Means – using the built training models is presented in table 7. Thus, as it 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 that has the largest number of data instances; a group, which has the least data instances (exceptions); a group that includes 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 respectively. As we can see, the



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

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 instance of various age groups (figure 12). Also, a small number of data instances of various age groups is present in Cluster 2, received in the course of implementation of the K-Means algorithm (figure 13). Cluster 0 and Cluster 2 formed with the Expectation Maximization algorithm as well as Cluster 1 and Cluster 2 formed with the Farthest First algorithm contain homogeneous age groups, and Cluster 0 and Cluster 1, formed with K-Means algorithm.



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

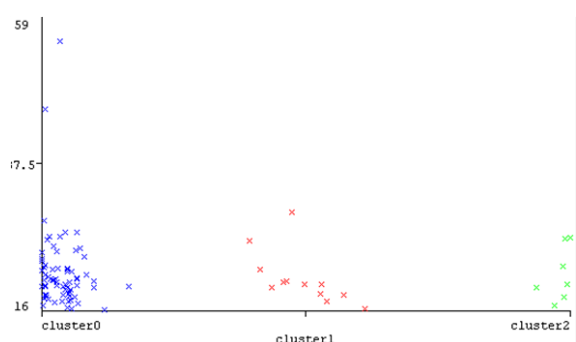


Figure 12: Plot of cluster distribution applying the Farthest First algorithm depending on the age group attribute.

Figures 14, 15 and 16 present a graphic representation

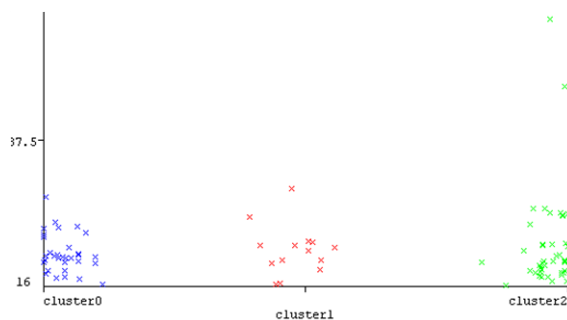


Figure 13: Plot of cluster distribution applying the K-Means algorithm depending on the age group attribute.

tation 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 visualizes clustering through application of the Expectation Maximization algorithm, shows that Cluster 0 contains only female data instances. Clusters 1 and 2 have data 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 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 both male and female data instances (figure 16).

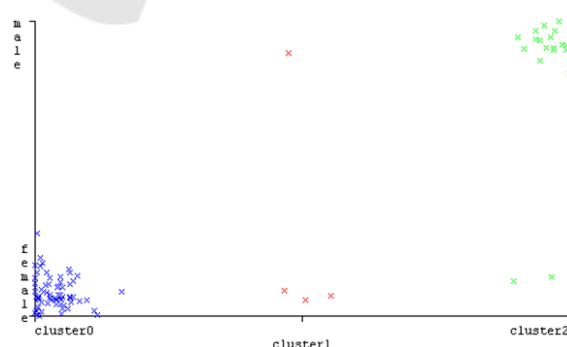


Figure 14: Plot of cluster distribution applying the Expectation Maximization algorithm depending on the sex attribute.

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

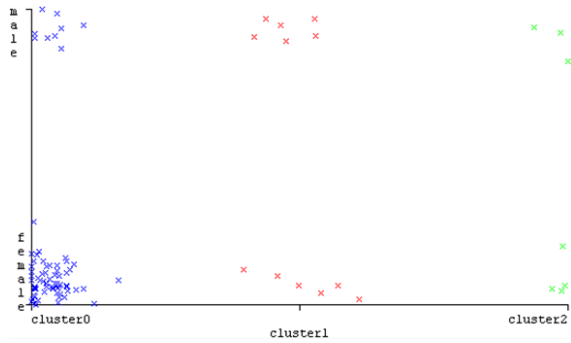


Figure 15: Plot of cluster distribution applying the Farthest First algorithm depending on the sex attribute.

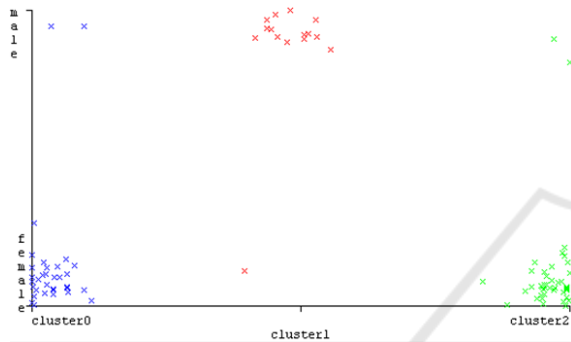


Figure 16: Plot of cluster distribution applying the K-Means algorithm depending on the sex attribute.

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

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 percent 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 a higher level of 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 the course of application of 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 application of the Vote (classifiers.rules.ZeroR) model.

According to the Mean absolute error (MAE), the data forecast that is closes to the actual results both in the process pf 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 0.3643 (testing – 0.3569) model. The approximately twice higher MAE value was received in course of 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 as a result of 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 prioritization 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 characterizes an almost random prediction.

## 4 CONCLUSION

In the course of determine the fields of use and conduct an empirical comparison of ensemble classification and clustering methods using the WEKA machine learning system to study the signs of IA related disorders of students, the following conclusions have been made:

1. As a result of empirical comparison of Expectation Maximization, Farthest First and K-Means algorithms using the WEKA machine learning system, we developed models of data instances' clustering to determine the signs of internet addiction disorders among students majoring in Computer Sciences.
2. The implementation of the Expectation Maximization, the K-Means and the Farthest First algorithms each resulted in the formation of 3 clusters. The results of clustering demonstrate that Internet centration in the psychic reality of a personality is a characteristic feature of the respondents that took part in the survey. This also reflects accordingly in their activity and behavior, diminishing other life interests and the significance of everyday activities. In addition, in the course of implementation of the Expectation Maximization al-

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%

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	

gorithm, a cluster was formed, whose representatives have behavior control disorders, related to online gaming. These respondents are in the risk group for developing IA related disorders.

- Expectation Maximization, Farthest First and K-Means algorithms of data clustering differ by their algorithm model, however, from the point of characteristic features, they produce relatively similar clusters, thus implementing optimized clustering. At the same time, when a data set was grouped into three clusters by implementing these algorithms, the clusters differed by cluster model, namely, by the number of data instances in each cluster, their structure and value of attribute centroids.
- Judging by the evaluation results of clustering validity using the validity indices, we can state that most likely the K-Means and Farthest First algo-

gorithms show worse clustering results than the Expectation Maximization algorithm.

- Respondents are divided into three groups (Significant Risk (SR), Insignificant Risk (IR), No Risk (NR)). Such division gives the possibility of primary general assessment of risks of IA development based on the significance of Internet influence on the psychic of a person. The Significant Risk (SR) group is determined by asking questions, which reflect the signs of “in-depth”, maladaptive and, accordingly, a relatively long-lasting influence of the Internet on the psychic, vital resources, vitality, the existential level, the personality in general in its vital and conceptual basis. The Insignificant Risk (IR) is determined by asking question, which disclose the signs of “superficial”, local, adaptive even though a rather significant influence on the psychic. In this group

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

		Actual class			
		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%

the spheres, influence by the Internet are the cognitive, activity, value-conceptual, need, communicative spheres of the psychic. The No Risk (NR) group indicates the absence of risks for IA development. Belongingness to this group is defined by asking questions, which reflect an insignificant, local and short-term influence of the Internet on the psychic.

- The model that gave the results that are the closest ones to the actual classification results is the model built using the Random Forest algorithm. According to all the assessments, the classification model built using the Bagging algorithm (classification algorithm classifiers.trees.REPTree) is close to it. Somewhat lower classification indexes are received in the course of building a model using the Ada Boost algorithm (classifiers.trees.DecisionStump). These models can be considered suitable for diagnosing IA disorders among students. The model built with the help of the Vote algorithm (classifiers.rules.ZeroR) is not suitable for use. Such a result indicates that the

application of this algorithm requires additional modifications.

- Intellectual analysis of the data set regarding the situation with IA among students majoring in Computer Sciences with the application of ensemble classification and clustering methods has shown that the methods studied above may be considered suitable for developing models for detecting IA disorders and respondent groups with the signs of IA related disorders.
- The results of the research indicate the expedience of the application of the intellectual data analysis in medical research using the machine learning systems. The presented methods may serve as the basis for a strategic development of new vectors of medical data processing as well as decision-making in this field.

The present-day medicine needs non-standard approaches to intellectual data analysis, complex application of methods, their modification, application the ensemble of methods in order to be able to process large data sets in digital systems. Our conclusions may help to determine the signs of IA related disorder.

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	

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

	Ensemble classification algorithm scheme	Actual class			
		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

ders among students majoring in Computer Sciences, forecasting the risk of IA and development of services aimed at IA prevention.

## REFERENCES

Abbott, D. A., Cramer, S. L., and Sherrets, S. D. (1995). Pathological gambling and the family: Practice implications. *Families in Society*, 76(4):213–219.

Anokhin, P. (1968). Cybernétique, neurophysiologie et psychologie. *Social Science Information*, 7(1):169–197.

Balatskiy, E. V. (2008). Vitalnyye resursy i kontury soznaniya (Vital resources and circuits of consciousness). *Vestnik Rossiyskoy akademii nauk*, 78(6):531–537.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

da Silva, L. E. B., Melton, N. M., and au2, D. C. W. I. (2019). Incremental cluster validity indices for hard partitions: Extensions and comparative study.

Dasgupta, S. and Long, P. M. (2005). Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4):555 – 569.

Derhach, M. (2016). Cyber-addiction of students majoring in computer science. *Science and Education*, (7):92–98.

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*, 10:100200.

Frankl, V. E. (1985). *Man’s search for meaning*. Simon and Schuster.



- Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, ICML'96*, page 148–156, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- 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):1233.
- Huizinga, J. (2016). *Homo Ludens: A Study of the Play-Element in Culture*. Angelico Press.
- 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):357–379.
- ICD-11 for Mortality and Morbidity Statistics (Version: 09/2020) (2020). 6C51 Gaming disorder. <https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1448597234>.
- Ji, H.-M., Chen, L.-Y., and Hsiao, T.-C. (2019). Real-time detection of internet addiction using reinforcement learning system. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1280–1288.
- Keng, B. (2016). The expectation-maximization algorithm. <http://bjlkeng.github.io/posts/the-expectation-maximization-algorithm>.
- Kittler, J., Hatef, M., Duin, R. P., and Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239.
- 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*, 2546:58–75.
- Klochko, O., Fedorets, V., Tkachenko, S., and Maliar, O. (2020). The use of digital technologies for flipped learning implementation. *CEUR Workshop Proceedings*, 2732:1233–1248.
- Klochko, O. V. (2019). *Matematychnye modeliuвання system i protsesiv v osviti/pedahohitsi (Mathematical modeling of systems and processes in education/pedagogy)*. Druk.
- 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):85–105.
- Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.
- Lee, T.-H., Ullah, A., and Wang, R. (2020). Bootstrap aggregating and random forest. In Fuleky, P., editor, *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*, pages 389–429. Springer International Publishing, Cham.
- Leontyev, D. A. (2017). *Psikhologiya smysla: priroda, stroeniye i dinamika smyslovoy realnosti (The psychology of meaning: nature, structure and dynamics of meaningful reality)*. Litres.
- Linoff, G. S. and Berry, M. J. A. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.
- 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):541–563.
- Pacol, C. A. and Palaog, T. D. (2020). Enhancing sentiment analysis of textual feedback in the student-faculty evaluation using machine learning techniques. In *3rd International Conference On Academic Research in Science, Technology and Engineering*. <https://www.dpublication.com/wp-content/uploads/2020/11/3-3003.pdf>.
- Santos, S. G. T. d. C. and de Barros, R. S. M. (2020). Online AdaBoost-based methods for multiclass problems. *Artificial Intelligence Review*, 53(2):1293–1322.
- 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):17111–17121.
- Subasi, A., Kevric, J., and Abdullah Canbaz, M. (2019). Epileptic seizure detection using hybrid machine learning methods. *Neural Computing and Applications*, 31(1):317–325.
- Sudakov, K. V. (2011). *Funktsionalnyye sistemy (Functional systems)*. Izdatelstvo RAMN.
- Suma, S. N., Nataraja, P., and Sharma, M. K. (2021). Internet addiction predictor: Applying machine learning in psychology. In *Advances in Artificial Intelligence and Data Engineering*, pages 471–481. Springer.
- Tarasenko, A., Yakimov, Y., and Soloviev, V. (2019). Convolutional neural networks for image classification. *CEUR Workshop Proceedings*, 2546:101–114.
- Tuysuzoglu, G. and Birant, D. (2020). Enhanced bagging (ebagging): A novel approach for ensemble learning. *The International Arab Journal of Information Technology*, 17(4):515–528.
- Wallis, D. (1997). Just click no: Talk story about Dr. Ivan K. Goldberg and the internet addiction disorder. *The New Yorker*. <http://www.newyorker.com/magazine/1997/01/13/just-click-no>.
- Weka (2021). Weka 3: Machine Learning Software in Java. <http://old-www.cms.waikato.ac.nz/ml/weka/>.
- Young, K. S. (1998a). *Caught in the net: How to recognize the signs of internet addiction—and a winning strategy for recovery*. John Wiley & Sons.
- Young, K. S. (1998b). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior*, 1(3):237–244. <https://www.liebertpub.com/doi/pdf/10.1089/cpb.1998.1.237>.
- Yuryeva, L. N. and Bolbot, T. Y. (2006). *Kompyuternaya zavisimost: formirovaniye, diagnostika, korrektsiya i profilaktika (Computer addiction: formation, diagnostics, correction)*.

- and prevention*). Porogi, Dnepropetrovsk. <http://kingmed.info/media/book/3/2673.pdf>.
- 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):012021.
- Zelinska, S. (2020). Machine learning: Technologies and potential application at mining companies. *E3S Web of Conferences*, 166:03007.
- Zhong, Y., Yang, H., Zhang, Y., and Li, P. (2020). Online random forests regression with memories. *Knowledge-Based Systems*, 201:106058.

