Електронний журнал «Ефективна економіка» включено до переліку наукових фахових видань України з питань економіки (Категорія «Б», Наказ Міністерства освіти і науки України № 975 від 11.07.2019). Спеціальності – 051, 071, 072, 073, 075, 076, 292.
Ефективна економіка. 2026. № 2.
ISSN 2307-2105

DOI: http://doi.org/10.32702/2307-2105.2026.2.158

УДК 330.1:342.5

V. Varenyk,

Postgraduate student of the Department of Economics, Finance and Accounting, Private Higher Education Establishment "European University"

ORCID ID: https://orcid.org/0009-0002-9777-4899

SUPERVISED MACHINE LEARNING ALGORITHMS FOR FINANCIAL MONITORING

В. В. Вареник,

аспірант кафедри економіки, фінансів та обліку,

Приватний вищий навчальний заклад «Європейський університет»

АЛГОРИТМИ КОНТРОЛЬОВАНОГО МАШИННОГО НАВЧАННЯ В ФІНАНСОВОМУ МОНІТОРИНГУ

Global digitalization of financial transactions, increase in transaction speed and number leads to impossibility of financial institutions to fulfill requirements in the field of financial monitoring without introduction of automated transaction verification. Therefore, the purpose of the article is to analyze the possibility of using supervised machine learning algorithms in automation systems of primary

financial monitoring entities. The use of machine learning will allow not only to optimize the internal resources of institutions, which are necessary for processing transactions, but also to increase the transparency of the distribution of financial transactions into classes of ordinary or risky transactions, which will have a positive impact on proving to state regulators the functioning of an effective risk management system in the institution in the field of combating money laundering and terrorist financing. The results of the study revealed that logistic regression, Naive Bayes and CART meet the requirements of reporting institutions in the field of financial monitoring regarding the efficiency of assessing new transactions, transparency of interpretation of the distribution of elements into classes and flexibility in the settings of machine learning algorithms. It has been established that the k-NN and SVM algorithms have limitations when used in automation systems, which are explained by the low speed of learning models based on these algorithms, the need for institutions to have significant technical resources to calculate the classification results, and the impossibility of visually interpreting the results of transaction classification. It has been proven that logistic regression and Naive Bayes are not able to detect complex nonlinear dependencies, therefore their use in information systems of financial institutions requires constant updating of the data set on which the model is trained, and the presence of an alternative classification tool aimed at identifying complex money laundering schemes. As an additional technique for monitoring complex financial transactions, k-NN and SVM algorithms can be used, since the specifics of their work allow detecting complex relationships between many features of elements. If a reporting institution has resource constraints in financial monitoring and needs to use only one supervised machine learning algorithm, the logical choice is to use a model based on the CART technique, which combines simplicity of calculation with the ability to search for complex nonlinear relationships.

Глобальна цифровізація фінансових операцій, зростання швидкості транзакцій та їх кількості призводить до неможливості виконання

фінансовими установами вимог у сфері фінансового моніторингу без запровадження автоматизованої перевірки транзакцій. Відтак, метою статті є аналіз можливості застосування алгоритмів контрольованого машинного навчання в системах автоматизації суб'єктів первинного фінансового моніторингу. Використання машинного навчання дозволить не лише оптимізувати внутрішні ресурси установ, які необхідні для обробки транзакцій, а й підвищить прозорість розподілу фінансових операцій на класи звичайних або ризикових транзакцій, що позитивно вплине на доведення державним регуляторам функціонування в установі ефективної системи управління ризиками в сфері протидії відмиванню коштів та фінансуванню тероризму. За результатами дослідження виявлено, що логістична регресія, Naive Bayes та CART відповідають вимогам звітних установ у сфері фінансового моніторингу щодо оперативності оцінки нових транзакцій, прозорості інтерпретації розподілу елементів на класи та гнучкості в налаштуваннях роботи алгоритмів машинного навчання. Встановлено, що алгоритми k-NN та SVM мають обмеження при використанні в системах автоматизації, які пояснюються низькою швидкістю навчання моделей, заснованих на даних алгоритмах, необхідністю наявності в установ значних технічних ресурсів для обчислення результатів класифікації та неможливістю наочної інтерпретації результатів класифікації транзакцій. Доведено, що логістична регресія та Naive Bayes не здатні виявляти складні нелінійні залежності, тому їх використання в інформаційних системах фінансових установ потребує постійного оновлення набору даних, на яких здійснюється навчання моделі, та наявністю альтернативного інструменту класифікації, спрямованого на ідентифікацію складних схем відмивання коштів. В якості додаткової техніки моніторингу складних фінансових транзакцій можуть бути використані алгоритми k-NN та SVM, оскільки специфіка їх роботи дозволяє виявляти складні взаємозв'язки між багатьма ознаками елементів. При наявності в звітної установи з фінансового моніторингу ресурсних

обмежень та необхідності використання лише одного алгоритму контрольованого машинного навчання, логічним вибором є використання моделі на основі техніки CART, яка поєднує простоту обчислення з можливістю пошуку складних нелінійних взаємозв'язків.

Keywords**:** *financial monitoring, supervised machine learning, risk, transaction analysis, primary financial monitoring entity, automation system.*

Ключові слова**:** *фінансовий моніторинг, контрольоване машинне навчання, ризик, аналіз транзакцій, суб'єкт первинного фінансового моніторингу, система автоматизації.*

Problem statement. Ongoing monitoring of business relationships and customer financial transactions, in accordance with national and international financial monitoring requirements, is an integral part of Customer Due Diligence measures. Reporting entities must analyze financial operations for consistency with available customer information and identify suspicious transactions that may indicate potential money laundering or terrorist financing attempts. Reporting institutions are obligated to document the measures taken regarding transaction monitoring by maintaining relevant electronic documents and records. This approach should facilitate the most rational and efficient use of the financial institution's resources and demonstrate to regulatory authorities that decisions based on transaction monitoring are evidence-based and stem from comprehensive analysis. Therefore, when selecting automation tools for financial transaction analysis, institutions should choose solutions that balance operational efficiency with the explainability of the decisions made. Models based on supervised machine learning algorithms meet these requirements. Consequently, prior to implementing these algorithms, it is essential to evaluate their core operating principles and the feasibility of their practical integration into the automation systems of reporting institutions.

Analysis of research and publications. Numerous researchers have dedicated their scientific works to the application of supervised machine learning algorithms for data classification. However, these studies have typically focused on fields unrelated to anti-money laundering (AML) or countering the financing of terrorism (CFT). Specifically, Zacharis N. Z. utilized the CART decision tree algorithm to predict student performance [13]. A research team led by Panda, N. R. analyzed a logistic regression model for application in the healthcare sector [10]. The study by Roy, A. and Chakraborty, S. addressed the use of Support Vector Machines (SVM) for structural reliability analysis [11]. The work of Xu, S. focused on measuring the performance of three variants of the Naive Bayes algorithm for text classification purposes [12]. Cherif, W. demonstrated the effectiveness of the k-NN algorithm for breast cancer diagnosis [5]. Consequently, the use of supervised machine learning algorithms for financial monitoring purposes remains insufficiently explored.

Formulation of the article's objectives. The aim of this article is to investigate the feasibility of applying supervised machine learning algorithms within the automation systems of primary financial monitoring entities to ensure compliance with requirements for analyzing financial transactions for potential money laundering or terrorist financing risks.

The paper main body. Logistic regression is a supervised machine learning algorithm that predicts the probability that an object belongs to a particular class. Despite its name, the algorithm is used for classification purposes.

At the first stage of calculating logistic regression [2, p. 3], a linear combination of features is calculated:

$$z = w_0 + w_1 x_1 + \omega_2 x_2 + \cdots + w_n x_n \quad (1)$$

where z – weighted sum of features;

x – feature specified in the dataset;

w – weight coefficient.

When used in financial monitoring automation systems, the product of the weight coefficients of all transaction risk indicators allows obtaining an overall risk scoring of a financial transaction.

In the second stage, a logistic (sigmoid) function is used to determine the probability coefficient:

$$P = \frac{1}{1+e^{-z}} \quad (2)$$

where P – probability coefficient;

e – Euler's number.

The use of Euler's number simplifies model training via gradient descent, which is employed to find the optimal weighting coefficients. If the weighted sum is a large positive number, e^{-z} tends toward zero, and the probability score approaches «1». In the context of transaction analysis for financial monitoring, such a result indicates a high risk of money laundering. Conversely, if the weighted sum is a large negative number, e^{-z} becomes significantly large, and the probability score approaches «0». In terms of financial monitoring, this result indicates a low risk of money laundering.

At the last stage of the calculation, the model receives a probability value of the object belonging to a certain class, which is in the range from «0» to «1». For AML/CFT, there will be only two classes – suspicious or ordinary financial transaction. The specific threshold above which a transaction will be identified as suspicious is set in the internal policies of a specific subject of primary financial monitoring, taking into account the risk appetite of the financial institution.

To train a model based on the logistic regression algorithm, the logistic loss function is used [2, p. 3]:

$$L(y, \hat{p}) = -[y\, ln(\hat{p}) + (1-y)\, ln(1-\hat{p})] \quad (3)$$

where L – the logistic loss;

y – actual class label from the dataset;

$\hat{p}$ – model prediction given by the sigmoid function.

To avoid overtraining the model, two regularization options are used: L1-regularization (Lasso), which reduces the values of insignificant features, or L2-regularization (Ridge), which reduces the weight coefficients but leaves them in the resulting part of the model.

A popular probabilistic supervised learning algorithm is Naive Bayes, which is based on Bayes' theorem. The algorithm is called "naive" because it assumes complete independence of the features being studied. The general Bayes formula has the following form [4, p. 3]:

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{\sum_{i=1}^{n} P(x|C_i)P(C_i)} \quad (4)$$

where $P(C_i|x)$ – posterior probability (i.e. the probability that a transaction is suspicious given the observation of x features);

$P(x|C_i)$ – probability of observing characteristics x given that the transaction is suspicious;

$P(C_i)$ – a priori class probability (i.e. the overall probability of a transaction being suspicious, based on analysis of historical data);

$P(x|C_i)P(C_i)$ – total probability of observing trait x in the entire sample.

Naive Bayes classification is based on the application of the maximum likelihood principle to classify an object into the most likely category. Since the denominator in formula (4) is the same for all classes, the overall probability of a suspicious transaction can be expressed as the product of the individual probabilities:

$$P(c_i|x) = Max\{P(c_1|x), P(c_2|x) \dots P(c_n|x)\} \quad (5)$$

For financial monitoring purposes, a model based on Naive Bayes multiplies the total frequency of suspicious transactions by the probability of suspicion of each feature in the analyzed transaction. The result obtained is compared with the probability that the transaction is not risky, and the higher probability is selected. Thus, the model classifies each analyzed financial transaction as suspicious or normal. A potential problem with the use of Naive Bayes for AML/CFT purposes

is the situation when a certain feature in the training set does not contain signs of risk, but transactions with such a feature are actually recognized as risky by employees of primary financial monitoring entities. In this case, the model will assign a risk probability of «0» to such a feature, which will mean setting a zero risk value for the transaction as a whole and classifying it as «normal» even in cases where the risk probability determined for other features of the transaction is high. To avoid the «zero probability» problem, Laplace Smoothing is used, which adds a small number to each probability calculation and does not change the overall classification of the financial transaction.

Machine learning algorithms based on decision trees predict the value of a target variable by learning decision rules. These rules, in turn, are derived from the features of the training data set. A node in the tree corresponds to a test for an attribute, a branch of the tree represents the test result, and a leaf of the tree contains the label of the predicted class. The prediction obtained from a decision tree is explained by a series of rules. The main components of any decision tree are a selection criterion and a stopping rule. For a categorical target variable, the selection criterion can be the classification error, the Gini index, or the information entropy. The Decision Tree searches among all predictor variables at all possible split points to find the optimal split that meets the given selection criterion. The algorithm repeats the split of objects recursively until the stopping rules take effect [14, p. 1050-1051].

A feature of the decision tree based on the CART (Classification and Regression Tree) algorithm is the generation of only two leaf nodes at each step of creating the tree. Accordingly, the objects of the study are divided into a part for which the rule is fulfilled and a part in which objects for which the rule is not fulfilled are collected. The selection of the optimal rule is achieved through the function of assessing the quality of the data set discrimination, which is based on the calculation of the Gini index. CART works on the principle of minimizing heterogeneity, therefore the algorithm for division selects the smallest weighted sum of the Gini index for both formed branches. For the purposes of financial

monitoring, CART divides transactions into the class of "regular transactions" and the class of "risky transactions" by analyzing all possible features of a financial transaction and all possible values of each feature. The result of the analysis is the determination of potential cut points. The algorithm selects the cut point at which the Gini index is minimal.

The Gini index formula is defined as follows [9, p. 544]:

$$Gini(D) = 1 - \sum_{j=1}^{n} P_j^2 \quad (6)$$

where $Gini(D)$ – Gini index for a data set D;

n – total number of classes in the dataset D;

P_j – proportion (probability) of objects of class j in a node.

When dividing the data set D into two subsets (a class of "normal" transactions and a class of "suspicious" transactions), the values $n1$ and $n2$ will represent the number of objects in subsets $D1$ and $D2$ respectively:

$$Gini(D) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2) \quad (7)$$

The specificity of the CART algorithm allows the model to detect complex nonlinear dependencies and combinations of features. Accordingly, the usefulness of the model for analyzing transactions for AML/CFT risks lies in the ability to identify typologies of money laundering, examples of which may not be explicitly contained in the training data set. In addition, the presentation of the analysis results in the form of a hierarchical rule structure provides a visual interpretation of the results, which is useful for proving the presence of an effective financial monitoring system to auditors or regulators.

Another popular machine learning algorithm for classification is called support vector machines (SVM). Using hyperplanes and kernel functions, the algorithm determines the optimal hyperplane for the distribution of two classes. SVM aims to form the best boundary or hyperplane that divides the n-dimensional space into different classes, placing new objects in the appropriate classes. The closest data points that affect the position of the hyperplane form support vectors.

The algorithm chooses the hyperplane with maximum margins, which maximizes the distance between two data points of different classes.

In the case of the possibility of dividing objects of two classes by a straight line, a linear type of SVM (or "hard margin SVM") is used. In this case, the data set for training the model, which consists of n-points, can be represented as the following formula:

$$\{(x_1 y_1), (x_2 y_2), \dots (x_n y_n)\} \quad (8)$$

Each point (x_i) – is a p – dimensional vector normalized by the values [-1, 1]. Accordingly, y_i takes the value «-1» or «1», depending on the class to which the point x_i belongs. The perpendicular to the hyperplane is denoted by «w». The parameter $\frac{b}{\|w\|}$ s equal in modulus to the distance from the hyperplane to the origin [5, p. 278].

Considering the above, the optimal distribution hyperplane can be described as:

$$w \cdot x + b = 0 \quad (9)$$

The formulas for the two support vectors will be as follows:

$$w \cdot x + b = 1 \quad (10)$$

$$w \cdot x + b = -1 \quad (11)$$

Since in "hard-margin SVM" the training data set can be linearly partitioned, the algorithm will choose hyperplanes such that there is no point from such training data set between them. After solving this problem, the algorithm maximizes the distance between hyperplanes. The width between support vectors is $\frac{2}{\|w\|}$, so for maximum performance of the SVM algorithm model, the value of $\|w\|$ needs to be minimized.

When classifying objects, errors are possible that will not allow for a linear distribution of data. The solution to this problem is to add attenuating variables to the objective function. The use of the SVM algorithm, in which the goal of the model is to correctly classify most objects, but not all objects, is called "SVM with

a soft margin". The value of the marginal error can be set by the financial monitoring entity independently, based on the risk appetite in the field of AML/CFT and the available resources for further processing of detected suspicious transactions.

Another problem of using the SVM algorithm is the possible presence of markers of different classes in the sample objects. For the correct classification of such elements, data mapping is used, in which, using kernel functions in the nonlinear type of SVM, objects are transformed into a higher-dimensional space in which there is a linear class separator. For mapping, the "kernel trick" method is usually used, the essence of which is the change of scalar products of points to a nonlinear kernel function, which is a scalar product of a higher-dimensional space [6, p. 283]. Kernel functions based on the "kernel trick" are a multilayer perceptron (sigmoid function), homogeneous and heterogeneous polynomials, Laplacian (exponential) function, spline, and Gaussian radial basis function. In addition, it is possible to use kernels based on Fourier series, B-splines, ANOVA splines, tensor products or additive kernels [1, p. 48].

The last machine learning algorithm that will be analyzed in this study is k-nearest neighbor (k-NN). The algorithm is used for both classification and regression purposes. When classifying, the method sequentially performs three main stages of analyzing a new object: it calculates the distances between the new object and the existing classified objects in the feature space, finds the nearest neighbors of the new object, and classifies the new object based on the belonging of the nearest objects to a certain class. For an existing data set, the algorithm predicts the correlation between new elements and existing elements of the model, and based on this correlation, classifies new elements into the category that best matches them. A feature of the k-NN algorithm is that the training process starts only at the moment when there is a need to classify new elements. Therefore, the training phase in k-NN includes only the storage of feature vectors and class labels for each data object. The k-nearest neighbor method is non-parametric, so in k-NN there is no predefined relationship between the input and output data [3, p. 3].

The main parameters of k-NN that affect the efficiency of the algorithm are the choice of the approach to calculating the distance between the model elements and the choice of the number of elements closest to the new object (parameter «k»). Typically, the distance from a new data element to each of the existing model elements is measured in the k-NN algorithm using the Euclidean distance.

The Euclidean distance between objects «p» and «q» by «n» features can be calculated using the following formula:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad (12)$$

The position of an object in n-dimensional Euclidean space is represented by a Euclidean vector. Therefore, «p» and «q» are Euclidean vectors that originate from the coordinates of the space, and the ends of the vectors define two points [8, p. 355].

In the case of a large difference in values between different types of explanatory variables, the Euclidean distance between the elements of the model will significantly depend on the features with larger ranges. To level this effect, the k-NN algorithm uses normalization. The main options for normalization are Z-normalization and minimax normalization [7, p. 16]. The minimax normalization formula leads to the fact that all values of the point «x» will be in the range from «0» to «1»:

$$x_{new} = \frac{x - min(x)}{max(x) - min(x)} \quad (13)$$

where x_{new} – new feature value obtained after normalization;

x – feature value before normalization;

$max(x)$ – maximum value in a data set;

$min(x)$ – minimum value in a data set.

When using Z-normalization, most of the normalized feature values will be in the range $(-3\sigma; 3\sigma)$. The normalization formula of this type has the following form:

$$x_{new} = \frac{x - \mu}{\sigma} \quad (14)$$

where x_{new} – new feature value obtained after normalization;

x – feature value before normalization;

μ – sample mean;

σ – standard deviation value.

The second critical element of the k-NN algorithm is setting the value of the parameter «k», which determines the number of elements closest to the object that already have a class label and that will be selected by the method to identify the class of a new data point. A large value of «k» reduces the influence of variance caused by random error. However, this creates the risk that the model will ignore small but important patterns. At the optimal value of «k», the algorithm finds a balance between underfitting and overfitting. To achieve this balance, in many k-NN studies, the parameter «k» is equal to the square root of the number of observations in the training data set [15, p. 2].

For the purposes of use in automated financial monitoring systems, the k-NN algorithm classifies a new transaction by comparing it with the «k» number of previously classified transactions that have the most similar feature parameters to the new transaction. Thus, the algorithm tries to determine how typical the financial transaction to be classified is. If a new object is in a space where there are no classified elements nearby or such elements have a «suspicious» label, the financial transaction is also classified as «suspicious».

Conclusions. The use of supervised machine learning algorithms in automation systems of primary financial monitoring entities depends on the efficiency of assessing a new transaction for the risk of money laundering or terrorist financing, the possibility of transparent interpretation of the assessment results, and the ability to configure the algorithm by reporting institutions in the field of AML/CFT, based on the overall risk appetite and resources of the financial institution.

The logistic regression forecast is based on the calculation of a linear combination of features and a sigmoid function. The model is trained using

gradient descent, which is used to identify the best weights for each AML/CFT risk feature. Identification of risky transactions using the Naive Bayes algorithm is carried out by multiplying the total frequency of suspicious transactions by the probability of suspicion of each feature in the analyzed transaction. The resulting product is compared with the probability that the transaction is not risky, after which the financial transaction is classified. The CART algorithm is a variation of the decision tree, in which binary classification is performed, and the distinction of the data set is based on the calculation of the Gini index. The above characteristics of logistic regression, Naive Bayes and CART meet the requirements for algorithms used by primary financial monitoring entities in automation systems. When using a single algorithm in an automation system, it is advisable to use CART, since in this technique the speed of calculation is combined with the identification of complex relationships between transaction features.

The k-NN algorithm is based on calculating the Euclidean distance from a new data element to each of the existing elements. The number of elements closest to the object is an important factor in the performance of the model and can be determined by calculating the square root of the number of observations in the training data set. The SVM algorithm is aimed at creating the best hyperplane that divides objects into classes. In the case of the presence of markers of different classes in the objects, data mapping is used, in which the nonlinear type of SVM uses kernel functions to transform objects into a higher-dimensional space in which there is a linear class separator. Accordingly, the use of k-NN and SVM in automated AML/CFT systems is limited by the speed of model training, computational complexity, and complexity of interpreting the results obtained.

Prospects for further research include the analysis of techniques for evaluating the performance of supervised machine learning algorithms and the study of the effectiveness of models based on supervised machine learning algorithms using specific data sets from the field of financial monitoring.

Література

1. Awad, M., Khanna, R. Support Vector Machines for Classification. *Efficient Learning Machines. Apress, Berkeley, CA*. 2015. DOI: https://doi.org/10.1007/978-1-4302-5990-9_3

2. Bailly, A., Blanc, C., Francis, É., Guillotin, T., Jamal, F., Wakim, B., Roy, P. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Computer Methods and Programs in Biomedicine*. 2022. № 213, 106504. DOI: https://doi.org/10.1016/j.cmpb.2021.106504

3. Bansal M., Goyal A., Choudhary A. A comparative analysis of K-Nearest Neighbour, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*. 2022. P. 100071. DOI: https://doi.org/10.1016/j.dajour.2022.100071

4. Chen, H., Hu, S., Hua, R., Zhao, X. Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*. 2021. № 2021 (30). DOI: https://doi.org/10.1186/s13634-021-00742-6

5. Cherif, W. Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis. *Procedia Computer Science*. 2018. № 127. P. 293-299. DOI: https://doi.org/10.1016/j.procs.2018.01.125

6. Cortes, C., Vapnik V. Support-vector networks. *Machine learning*. 1995. № 20. P. 273-297. DOI: https://doi.org/10.1007/BF00994018

7. Henderi, H., Wahyuningsih, T., Rahwanto, E. Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *International Journal of Informatics and Information Systems*. 2021. № 4(1). P. 13-20. DOI: https://doi.org/10.47738/ijiis.v4i1.73

8. Kataria, A., Singh, M. D. A review of data classification using k-nearest neighbour algorithm. *International Journal of Emerging Technology and*

Advanced Engineering. 2013. № 3(6). P. 354-360. URL: https://www.ijetae.com/Volume3Issue6.html (дата звернення 28.01.2026).

9. Lin, C. L., Fan, C. L. Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. *Journal of Asian Architecture and Building Engineering*. 2019. № 18(6). P. 539-553. DOI: https://doi.org/10.1080/13467581.2019.1696203

10. Panda, N. R., Pati, J. K., Mohanty, J. N., Bhuyan, R. A review on logistic regression in medical research. *National Journal of Community Medicine*. 2022. № 13(04). P. 265-270. DOI: https://doi.org/10.55489/njcm.134202222

11. Roy, A., Chakraborty, S. Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety*. 2023. № 233. P. 109126. DOI: https://doi.org/10.1016/j.ress.2023.109126

12. Xu, S. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*. 2018. № 44(1). P. 48-59. DOI: https://doi.org/10.1177/016555151667794

13. Zacharis, N. Z. Classification and regression trees (CART) for predictive modeling in blended learning. *International Journal of Intelligent Systems and Applications*. 2018. № 10(3). P. 1-9. DOI: https://doi.org/10.5815/ijisa.2018.03.01

14. Zhang Y., Trubey P. Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection. *Computational Economics*. 2018. № 54(3). P. 1043–1063. DOI: https://doi.org/10.1007/s10614-018-9864-z

15. Zhang Z. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*. 2016. № 4(11). P. 218. DOI: https://doi.org/10.21037/atm.2016.03.37

References

1. Awad, M. and Khanna, R. (2015), “Support Vector Machines for Classification”, *Efficient Learning Machines. Apress, Berkeley, CA*. DOI: https://doi.org/10.1007/978-1-4302-5990-9_3

2. Bailly, A. Blanc, C. Francis, É. Guillotin, T. Jamal, F. Wakim, B. and Roy, P. (2022), "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models", *Computer Methods and Programs in Biomedicine*, vol. 213. DOI: https://doi.org/10.1016/j.cmpb.2021.106504

3. Bansal M. Goyal A. and Choudhary A. (2022), "A comparative analysis of K-Nearest Neighbour, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning", *Decision Analytics Journal*, vol. 3. DOI: https://doi.org/10.1016/j.dajour.2022.100071

4. Chen, H. Hu, S. Hua, R. and Zhao, X. (2021), "Improved naive Bayes classification algorithm for traffic risk management", *EURASIP Journal on Advances in Signal Processing*, vol. 2021. DOI: https://doi.org/10.1186/s13634-021-00742-6

5. Cherif, W. (2018), "Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis", *Procedia Computer Science*, vol. 127, pp. 293–299. DOI: https://doi.org/10.1016/j.procs.2018.01.125

6. Cortes, C. and Vapnik V. (1995), "Support-vector networks", *Machine learning*, vol. 20, pp. 273–297. DOI: https://doi.org/10.1007/BF00994018

7. Henderi, H. Wahyuningsih, T. and Rahwanto, E. (2021), "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer", *International Journal of Informatics and Information Systems*, vol. 4, pp. 13–20. DOI: https://doi.org/10.47738/ijiis.v4i1.73

8. Kataria, A. and Singh, M.D. (2013), "A review of data classification using k-nearest neighbour algorithm", *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, pp. 354–360, available at: https://www.ijetae.com/Volume3Issue6.html (Accessed 28 Jan 2026).

9. Lin, C.L. and Fan, C.L. (2019), "Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan", *Journal of*

Asian Architecture and Building Engineering, vol. 18, pp. 539–553. DOI: https://doi.org/10.1080/13467581.2019.1696203

10. Panda, N.R. Pati, J.K. Mohanty, J.N. and Bhuyan, R. (2022), "A review on logistic regression in medical research", *National Journal of Community Medicine*, vol. 13, pp. 265–270. DOI: https://doi.org/10.55489/njcm.134202222

11. Roy, A. and Chakraborty, S. (2023), "Support vector machine in structural reliability analysis: A review", *Reliability Engineering & System Safety*, vol. 233. DOI: https://doi.org/10.1016/j.ress.2023.109126

12. Xu, S. (2018), "Bayesian Naïve Bayes classifiers to text classification", *Journal of Information Science*, vol. 44, pp. 48–59. DOI: https://doi.org/10.1177/016555151667794

13. Zacharis, N.Z. (2018), "Classification and regression trees (CART) for predictive modeling in blended learning", *International Journal of Intelligent Systems and Applications*, vol. 10, pp. 1–9. DOI: https://doi.org/10.5815/ijisa.2018.03.01

14. Zhang Y. and Trubey P. (2018), "Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection", *Computational Economics*, vol. 54, pp. 1043–1063. DOI: https://doi.org/10.1007/s10614-018-9864-z

15. Zhang Z. (2016), "Introduction to machine learning: k-nearest neighbors", *Annals of Translational Medicine*, vol. 4. DOI: https://doi.org/10.21037/atm.2016.03.37

Отримано редакцією журналу / Received: 04.02.26
Прорецензовано / Revised: 10.02.26
Схвалено до друку / Accepted: 19.02.26