УДК 004.9

О. М. Козачко, к. т. н., доц.; С. О. Жуков, к. т. н., доц.; Т. Є. Вуж, к. т. н., доц.; А. О. Лотоцький

МЕТОД АНАЛІЗУ РІВНЯ ЗНАНЬ ІНОЗЕМНОЇ МОВИ СТУДЕНТІВ ЗАКЛАДУ ВИЩОЇ ОСВІТИ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

Стаття присвячена розробці методу аналізу рівня знань іноземної мови студентів закладу вищої освіти на основі методів машинного навчання. За допомогою методів машинного навчання можна виявити закономірності та тенденції, що дозволять підняти рівень знань студентів закладів вищої освіти. Завданнями роботи відповідно до поставленої мети є різносторонній аналіз даних: аналіз предметної галузі, розвідувальний аналіз, побудова моделі об'єкта та визначення дисциплін, що впливають на рівень вивчення іноземної мови. Для пошуку виду математичної моделі класифікатора в роботі використані дерева рішень.

Для розробки методу аналізу рівня знань іноземної мови студентів закладу вищої освіти в роботі використано технології машинного навчання, за допомогою яких розроблено різні моделі дерев рішень з подальшим вибором найкращої з них. Виявлення закономірностей здійснюється за рахунок побудови дерев рішень на вибірках отриманих оцінок студентами Вінницького національного медичного університету ім. М. І. Пирогова у 2-му, 4-му та 6-му семестрах навчання. В рамках цієї роботи пошук виду класифікатора здійснюється на основі градієнтного бустингу та логістичної регресії. Проведені експерименти показали, що правила отримані за допомогою регресійної моделі точніше прогнозують рівень знань іноземної мови. На основі цих досліджень зроблені адекватні і достатньо точні висновки про виявлені закономірності.

Запропонований метод дозволяє виявити закономірності визначення рівня знань іноземної мови студентів закладів вищої освіти, використовуючи методи машинного навчання та забезпечує виявлення дисциплін, вивчення яких найбільше впливає на рівень знань іноземної мови. Для визначення рівня знань іноземної мови студентами закладу вищої освіти, було розроблено програмний модуль у вигляді веб-системи з використанням основних веб-технологій, який дозволяє автоматизовано розв'язати поставлену задачу з наданням рекомендацій щодо покращення рівня знань іноземної мови. Програмний модуль включає в себе веб-сайт з підключеною базою даних. Результати дослідження можуть бути ефективно використані для покращення сучасного навчального процесу.

Ключові слова: *дерева рішень, розвідувальний аналіз, побудова моделей, виявлення закономірностей, фактор впливу, Python.*

Вступ

Зі стрімким розповсюдженням інформаційних технологій збільшується і кількість збереженої інформації. Ця інформація може бути використана для проведення інформаційного аналізу виявлення загальних тенденцій та ознак, що впливають на них. За допомогою виявлених тенденцій та ознак можна оптимізувати роботу системи, підвищити продуктивність, скоротити кількість втрат. Для будь-кого є очевидною важливість вивчення англійської мови в наш час. Англійська мова відкриває нескінченні можливості у повсякденному та професійному житті, тому проблема є актуальною.

Враховуючи виклики та проблеми, з якими стикається сучасний навчальний процес, підвищується рівень використання сучасних інтелектуальних систем та алгоритмів для підвищення рівня освіти та викладання у навчальних закладах. Існує велика кількість досліджень, спрямованих на визначення закономірностей у цій сфері. Ці дослідження можуть бути ефективно використані для з'ясування та виявлення сучасних освітніх проблем, а також індивідуальних та колективних особливостей учнів та студентів, за допомогою впровадження процесу класифікації та регресійного аналізу набору даних.

Розглянемо кілька проблем та підходів до їх вирішення, які допоможуть визначити напрям ведення дослідження та запобігти виникненню цих та подібних проблем.

У роботі [1] описані та продемонстровані результати, отримані від використання алгоритмів аналізу даних. Визначено основні особливості, залежності та методи виділення

основних ознак та факторів із набору даних. Прогнозування характеристик студентів допоможе поділити їх на різні класи, що дозволить цим студентам розвивати комунікативні, лідерські навички та навички самоврядування під час навчання в університеті чи іншому закладі освіти. Результати показують, що оцінювання показників ефективності є невід'ємною частиною покращення сучасного навчального процесу.

Наступним прикладом є стаття [2], в якій, у свою чергу, розглянуто та систематизовано багато прикладів та ідей з інших джерел. Обчислювальне мислення являє собою термінологію, яка охоплює складний набір процесів міркувань, які проводяться для постановки проблеми та вирішення за допомогою обчислювального інструменту. Здатність систематизувати задачі та розв'язувати їх цими засобами наразі вважається навиком, який повинні розвивати всі студенти разом із мовою, математикою та іншими науками. Враховуючи, що інформатика має багато коренів у галузі математики, доцільно розмірковувати, чи можна впливати на навчання математики, пропонуючи студентам заходи, пов'язані з обчислювальним мисленням. У цьому сенсі у цій статті представлено систематичний огляд літератури щодо оприлюднених доказів навчання математики в діяльності, спрямованій на розвиток навичок обчислювального мислення. Проаналізовано сорок дві статті, в яких були представлені рішення для оцінки результатів навчання, опубліковані з 2006 по 2017 рік.

У статті [3] розглянуто питання, що є спільного у розвиненні комп'ютерного мислення та вивченні англійської мови, які методи є найбільш ефективними та простими. Ефективне навчання обчислювального мислення для учнів, що вивчають англійську мову корелюється з іншими формами контентного навчання. Аналіз комп'ютерного коду можна використовувати для побудови мета-усвідомлення обчислювальної семіотики, а візуальна природа деяких мов програмування, таких як Scratch, може сприяти розвитку грамотності. Найголовніше, що проекти, пов'язані з обчислювальним мисленням, чи то в створенні розповідей, чи в розробці електронних текстових проєктів, дають студентам широкі можливості висловити та розвинути власну ідентичність. Зараз формується вся сфера обчислювального мислення в освіті і важливо, щоб те, як ми навчаємо обчислювальному мисленню, найкраще відповідало потребам наших різних студентів.

Актуальність роботи полягає в тому, що за допомогою методів машинного навчання можна виявити закономірності та тенденції, що дозволять підняти рівень знань студентів закладів вищої освіти. Тому дослідження в цій галузі є актуальними.

Об'єктом дослідження є процес аналізу рівня знань іноземної мови студентів Вінницького національного медичного університету ім. М. І. Пирогова.

Предметом дослідження є методи машинного навчання іноземної мови студентів вищих навчальних закладів.

Метою цієї роботи є виявлення дисциплін, які найбільше впливають на рівень знань іноземної мови.

Завданнями роботи відповідно до поставленої мети є різносторонній аналіз даних: аналіз предметної галузі, розвідувальний аналіз, побудова моделі об'єкта та визначення дисциплін, що впливають на рівень вивчення іноземної мови. На основі цих досліджень можна буде зробити адекватні і достатньо точні висновки про виявлені закономірності.

Огляд методів прогнозування та постановка задачі

Для розробки методу аналізу рівня знань іноземної мови студентів закладу вищої освіти в роботі використано технології машинного навчання, за допомогою яких розроблено різні моделі дерев рішень з подальшим вибором найкращої з них. Початковими даними є оцінки з різних дисциплін, які представлені у 200-бальній шкалі. Припустимо Y – це оцінка студента з іноземної мови, а $X_1, X_2, \ldots, X_n$ – це оцінки різних дисциплін, які отримав студент протягом

семестру Тоді закономірність між оцінкою з іноземної мови та оцінками інших дисциплін будемо шукати у вигляді такого співвідношення:

$$Y = f(X_1, X_2, ..., X_n), \quad (1)$$

де n – кількість дисциплін.

Оцінку з іноземної мови розіб'ємо на чотири рівня: «незадовільно», «задовільно», «добре» та «відмінно». Тоді співвідношення (1) можна розглядати як класифікатор, який для вхідного вектору оцінок X встановлює відповідність рівню знань з іноземної мови.

Для пошуку виду математичної моделі класифікатора (1) в роботі використано дерева рішень, що будуються за такими методами: бустинг, багінг, стекінг тощо. Бустинг – це ансамблевий мета-алгоритм машинного навчання передусім для зменшення зсуву а також і дисперсії у навчанні з учителем, та сімейство алгоритмів машинного навчання, які перетворюють слабких учнів на сильних [4]. Багінг – це мета-алгоритм композиційного навчання, призначений для поліпшення стабільності і точності алгоритмів машинного навчання, що використовуються в статистичній класифікації та регресії [5]. Стекінг – один з найпопулярніших способів ансамблювання алгоритмів, тобто використання декількох алгоритмів для вирішення однієї з задач машинного навчання [6].

Для автоматизації вищезгаданих методів існують три найбільш потужні бібліотеки: Xgboost, Catboost і LightGBM. Враховуючи великий розмір датасету, велику кількість ознак та обмежені обчислювальні можливості, пропонується застосовувати бібліотеку LightGBM [7].

Аналіз вхідних даних моделей

Вхідними даними для виявлення закономірностей рівня знань іноземної мови студентів закладу вищої освіти є набір оцінок, що отриманні студентами Вінницького національного медичного університету за дисципліни 2-го, 4-го, 6-го семестрів навчання та результати першого етапу єдиний державний кваліфікаційний іспит (ЄДКІ). Інформація про дисципліни, на основі яких виявляються закономірності впливу на рівень знань іноземної мови наведено в таблиці 1.

Таблиця 1

Інформація про дисципліни, що впливають на рівень знань іноземної мови

№	Позначення	Назва дисципліни	Мінімальне значення	Максимальне значення	Середнє значення
1	2	3	4	5	6
1	X_1	Латинська мова та медична термінологія	122	200	167,6
2	X_2	Медична біологія	124	200	158,3
3	X_3	Медична та біологічна фізика	122	200	166,4
4	X_4	Основи економічних теорій	153	200	179,6
5	X_5	Основи психології та педагогіки	120	197	158,7
6	X_6	Безпека життєдіяльності та охорона праці	132	200	173,0
7	X_7	Біологічна та біоорганічна хімія	123	198	151,8
8	X_8	Догляд за хворими	142	200	177,4
9	X_9	Логіка, формальна логіка	128	196	163,1
10	X_{10}	Медична інформатика	131	195	165,4
11	X_{11}	Фізичне виховання	135	200	171,1
12	X_{12}	Іноземна мова	122	200	157,6
13	X_{13}	Цивільний захист	122	195	154,5
14	X_{14}	Військова гігієна	122	200	164,2
15	X_{15}	Загальна хірургія	131	200	170,7
16	X_{16}	Патоморфологія	99	200	156,0
17	X_{17}	Патофізіологія	92	195	152,8

Таблиця 1

1	2	3	4	5	6
18	X_{18}	Пропедевтика внутрішньої медицини	122	200	159,0
19	X_{19}	Пропедевтика педіатрії	122	200	162,2
20	X_{20}	Сестринська практика	150	200	179,5
21	X_{21}	Фармакологія	125	200	158,5

Для проведення розвідувального аналізу оцінок кожен семестр розглядався окремо. На рис. 1 побудовано графік кореляції для першого датасету (теплова кореляційна карта).

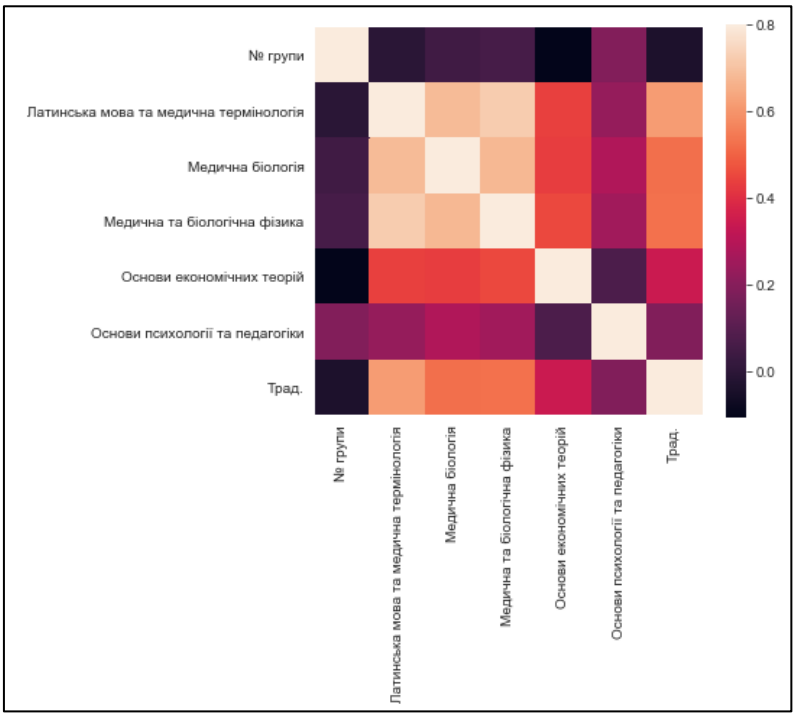

Рис. 1. Теплова кореляційна карта

На основі цієї карти можна зробити висновок, що існує деяка залежність між першими трьома дисциплінами. Але нас цікавить, що впливає на результати з англійської мови. Найвищу кореляцію з усіх доступних предметів має «Латинська мова та медична термінологія», що є логічним – здатність до вивчення одної мови впливає на вивчення інших мов. На рис. 2 наведено графіки розподілу значень по усіх дисциплінах.

На основі графіків розподілу можна зробити висновок, що найбільш схожим до розподілу дисципліни «Іноземна мова» є значення оцінок «Латинська мова та медична термінологія». Аналогічним чином було проаналізовано інші датасети.

```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Іноземна мова')
sns.violinplot(y='Іноземна мова',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Латинська мова та медична термінологія')
sns.violinplot(y='Латинська мова та медична термінологія',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Медична біологія ')
sns.violinplot(y='Медична біологія',data=small,palette='spring',linewidth=3)
plt.show()
```

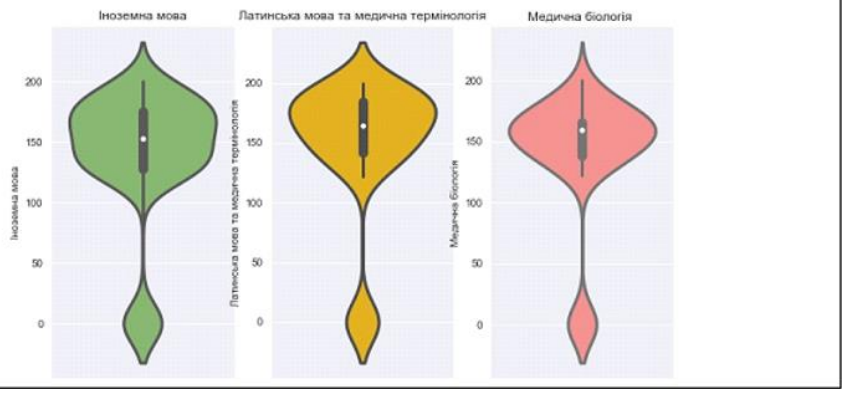

Рис. 2. Графіки розподілу значень перших трьох дисциплін

Розробка моделі об'єкта

В рамках цієї роботи пошук вигляду класифікатора (1) здійснюється на основі градієнтного бустингу [3] та логістичної регресії [4].

Задачу побудови класифікатора сформулюємо таким чином. Дано множину об'єктів оцінок X і їх класи $Y=\{2,3,4,5\}$, а також цільова функція y^*, значення якої $y_i = y^*(x)$ відомі на деякій підмножині класів $\{x_i\}$ множини X, де x_i – вектор розмірності n. Необхідно побудувати таку функцію, яка по відомим даним (x_i, y_i) наближає цільову функцію y на всій вибірці даних, $i = 1,..m$.

Для побудови класифікатора розглянемо градієнтний бустинг. Ідея градієнтного бустингу полягає в побудові ансамблю дерев рішень таким чином, щоб кожне наступне дерево намагалося покращити якість всієї комбінації дерев.

Класифікація здійснюється за наступною формулою [8]:

$$obj\_boost(X) = \underset{y \in Y}{argmax} \sum_{k:b_k(x)=y}^{n} a_k, \tag{2}$$

де $b_k(x)$ – відповідь k-го дерева на об'єкт x; a_k – вклад k-го дерева в композицію.

В процесі навчання послідовно будуються K дерев рішень на всіх m об'єктах та s випадково обраних ознак з загальної кількості ознак n. Після навчання чергового дерева, ваги невірно класифікованих об'єктів зростають, тим самим наступне дерево здійснює фокусування в основному на них.

Іншим класифікатором, який розглядається в роботі, є логістична регресія, яка представляє собою статистичну лінійну модель класифікації, що дозволяє спрогнозувати апостеріорні ймовірності класів за допомогою логістичної кривої. Об'єкт відноситься до класу з найбільшою ймовірністю, що визначається за такою формулою [8]:

$$obj\_reg(X) = \underset{y \in Y}{argmax} P(y^*(x) = y), \tag{3}$$

$$P(y^*(x) = y) = \frac{e^{<x,a_y>}}{\sum_{k=1}^{K} e^{<x,a_k>}}, \tag{4}$$

тобто об'єкту x присвоюється клас з найбільшою ймовірністю, яка обчислюється згідно softmax-функції a_k – вектор регресійних коефіцієнтів, які пов'язані з класом k, а $x = (x_1, \ldots x_n)$ – вектор ознак.

Дослідження показали, що логістична регресія є найкращою моделлю класифікації. На рис. 3 – 5 зображено дерева рішень, які отримані за допомогою логістичної регресії (3) на датасетах результатів навчання 2-го, 4-го та 6-го семестру навчання відповідно. З побудованих дерев рішень видно, що основними дисциплінами, які впливають на рівень знань іноземної мови є такі як «Латинська мова та медична термінологія», «Біологічна та біоорганічна хімія» та «Загальна хірургія». Аналогічним чином моделюється датасет даних за результатами ЄДКІ.

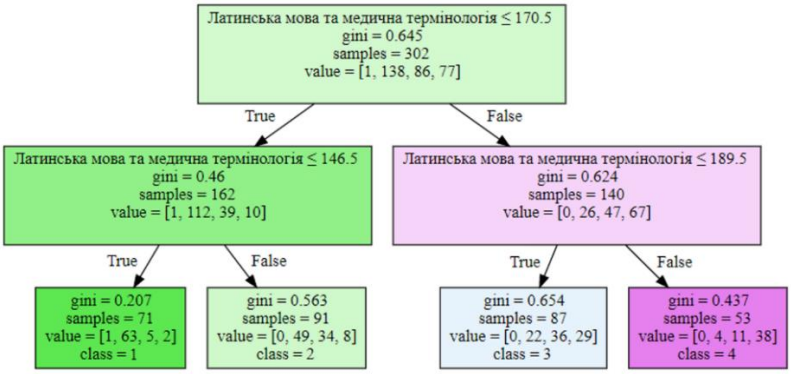

Рис. 3. Дерево рішень першого датасету за даними другого семестру

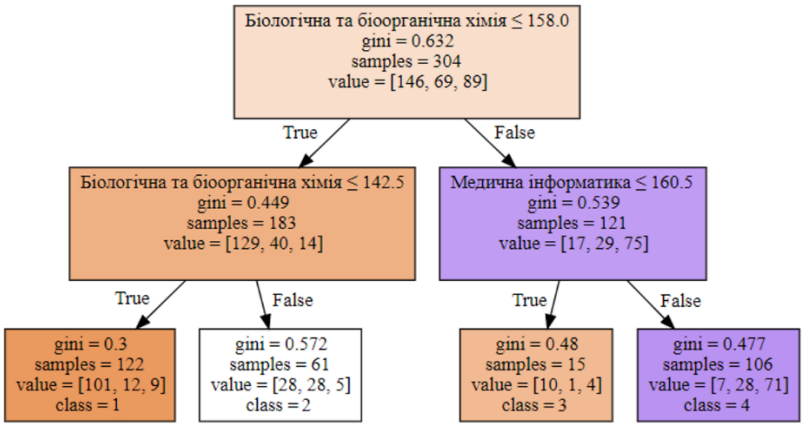

Рис. 4. Дерево рішень другого датасету за даними четвертого семестру

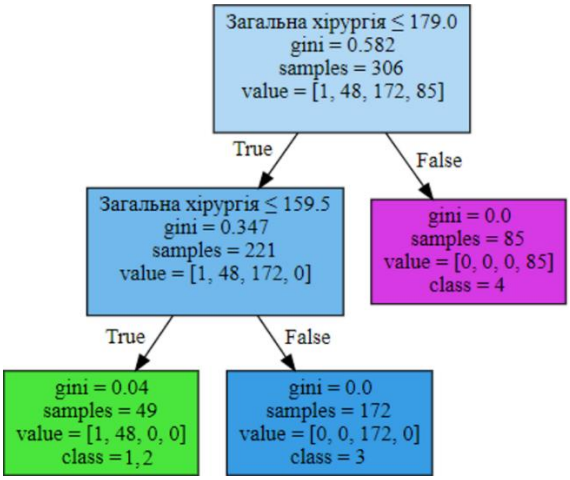

Рис. 5. Дерево рішень третього датасету за даними шостого семестру

З рис. 3 видно закономірність, яку можна інтерпретувати у вигляді такого правила «Якщо-То»:

– якщо студент отримав оцінку з дисципліни «Латинська мова та медична термінологія» менше 146.5 балів, то оцінка з іноземної мови буде «незадовільно»;

– якщо оцінка з дисципліни «Латинська мова та медична термінологія» в діапазоні від 146,5 балів до 170.5, то оцінка з іноземної мови буде «задовільно»;

– якщо оцінка з дисципліни «Латинська мова та медична термінологія» в діапазоні від 170,5 балів до 189.5, то оцінка з іноземної мови буде «добре»;

– якщо оцінка з дисципліни «Латинська мова та медична термінологія» більше 170.5, то оцінка з іноземної мови буде «відмінно».

Аналогічно інтерпретуються дерева рішень, зображені на рис. 4, 5.

Точність класифікації моделей машинного навчання наведена в таблиці 2.

Таблиця 2

Точність класифікації моделей машинного навчання

	2 семестр	4 семестр	6 семестр	ЄДКІ
XGBoost	71%	65%	77%	60%
Logistic regression	84%	73%	99%	78%

Точність класифікації визначено за частотою помилки за формулами:

$$F(X) = \sum_{j=1}^{m} \frac{\Delta_j}{m}, \tag{5}$$

$$F(X) = \begin{cases} 1, \text{ if } obj_j(X) = Y_j, \\ 0, \text{ if } obj_j(X) \neq Y_j. \end{cases} \tag{6}$$

Для визначення рівня знань іноземної мови студентами закладу вищої освіти, було розроблено програмний модуль у вигляді веб-системи з використанням основних веб-технологій: мови гіпертекстової розмітки HTML, каскадних таблиць CSS та шаблонів бібліотеки Bootstrap для побудови макету сайту, локального серверу Denwer, веб-інтерфейсу phpMyAdmin, для роботи з базою даних, яка написана мовою запитів SQL. Також, було використано PHP для підключення макету сайту з базою даних. Програмний модуль включає в себе веб-сайт з підключеною базою даних.

Висновки

В роботі запропоновано метод виявлення закономірностей рівня знань іноземної мови студентів закладів вищої освіти, який на відміну від наявних, використовує методи машинного навчання та забезпечує виявлення дисциплін, вивчення яких найбільше впливає на рівень знань іноземної мови. Виявлення закономірностей здійснено за рахунок побудови дерев рішень на вибірках отриманих оцінок студентами Вінницького національного медичного університету ім. М. І. Пирогова у 2-му, 4-му та 6-му семестрах навчання. Дерева рішень будувалися за допомогою методів бустинга та логістичної регресії. Проведені експерименти показали, що правила, які отриманні регресійною моделлю точніше прогнозують рівень знань іноземної мови. Крім того виявлено дисципліни, що впливають на підвищення рівня знань з іноземної мови, а саме: «Латинська мова та медична термінологія» є найбільш впливовою дисципліною у 2-му семестрі, «Біологічна та біоорганічна хімія» в 4-му семестрі і «Загальна хірургія» в 6-му семестрі.

СПИСОК ЛІТЕРАТУРИ

1. Predicting Pupil's Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods [Електронний ресурс]. Режим доступу: https://link.springer.com/chapter/10.1007/978-3-030-16621-2_58.

2. Mathematics Learning through Computational Thinking Activities: A Systematic Literature Review

[Електронний ресурс]. Режим доступу: http://jucs.org/jucs_24_7/mathematics_learning_through_computational/jucs_24_07_0815_0845_barcelos.pdf.

3. Teaching computational thinking to english learners [Електронний ресурс]. Режим доступу: https://par.nsf.gov/servlets/purl/10073683.

4. Zhi-Hua Z. Ensemble Methods: Foundations and Algorithms / Zhou Zhi-Hua. – New York : Chapman and Hall/CRC, 2012. – 236 с.

5. Shinde A. Preimages for Variation Patterns from Kernel PCA and Bagging / A.Shinde, A. Sahu, D. Apley, G. Runger. // IIE Transactions. – 2014. – Vol. 46. – P. 429 – 456. DOI: 10.1080/0740817X.2013.849836.

6. A Kaggler's Guide to Model Stacking in Practice [Електронний ресурс]. Режим доступу: http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to- model-stacking-in-practice/.

7. Random classification noise defeats all convex potential boosters [Електронний ресурс] / Philip M. Long, Rocco A. Servedio // Machine Learning. – 2010. – 78. – P. 287 – 304. DOI: https://doi.org/10.1007/s10994-009-5165-z.

8. Module pandas_profiling. [Електронний ресурс]. Режим доступу: https://pandas-profiling.ydata.ai/docs/master/.

9. Наказ МОЗ України від 22.01.2021 №106 [Електронний ресурс]. Режим доступу: https://zakon.rada.gov.ua/laws/show/z0269-21#n7.

10. Some Studies in Machine Learning Using the Game of Checkers [Електронний ресурс]. Режим доступу: https://ieeexplore.ieee.org/document/5392560.

Стаття надійшла до редакції 20.12.2021.
Стаття пройшла рецензування 25.12.2021.

Козачко Олексій Миколайович – к. т. н., доцент, доцент кафедри системного аналізу та інформаційних технологій.

Жуков Сергій Олександрович – к. т. н., доцент кафедри системного аналізу та інформаційних технологій.

Вінницький національний технічний університет.

Вуж Тетяна Євгенівна – к. т. н., доцент кафедри біологічної фізики, медичної апаратури та інформатики.

Вінницький національний медичний університет імені М. І. Пирогова.

Лотоцький Андрій Олександрович – магістрант кафедри системного аналізу та інформаційних технологій.

Вінницький національний технічний університет.