

К.Ю. Островська

## ДОСЛІДЖЕННЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ОБРОБКИ РЕЗУЛЬТАТІВ ТЕСТУВАННЯ

*Анотація. Робота присвячена дослідженню методів інтелектуального аналізу даних для обробки результатів тестування з використанням мови Python.*

*Метою даної роботи є дослідження методів інтелектуального аналізу даних засобами мови Python.*

*Для досягнення поставленої мети необхідно вирішення наступних завдань:*

- дослідження етапів інтелектуального аналізу даних, особливостей підготовки і вивчення даних;
- дослідження задач інтелектуального аналізу даних і методів їх рішення;
- дослідження засобів мови Python для вирішення завдань інтелектуального аналізу даних;
- практичне рішення задач ІАД: побудова моделей інтелектуального аналізу даних засобами мови Python;
- аналіз якості побудованих моделей.

*У даній роботі був проведений інтелектуальний аналіз даних, отриманих з системи управління навчанням Moodle.*

*Ключові слова: інтелектуальний аналіз даних, кластеризації, прогнозування, система управління навчанням, Python.*

**Постановка задачі.** Інтелектуальний аналіз даних або Data Mining є однією з активно що розвиваються областей інформаційних технологій, призначеної для виявлення прихованих і нетривіальних знань з даних. Технології інтелектуального аналізу даних знаходять своє застосування в різних областях діяльності людини: наукові дослідження, охорону здоров'я, ділова галузь, фінансова сфера та багато інших.

Одним з таких напрямків є сфера освіти. На даний момент слід відзначити існування таких спеціальних систем в сфері освіти, які називаються «Системи управління навчанням».

Система управління навчанням (Learning Management System, LMS) - система для управління навчальною діяльністю, яка використовується для створення, управління і поширення навчальних матеріалів з забезпеченням доступу до інформації.

Система управління навчанням дозволяє:

- розміщувати навчальні матеріали, забезпечуючи доступ до них і розподіляючи навантаження;
- контролювати процес вивчення матеріалу і терміни виконання завдань;
- організовувати взаємодію учасників навчального процесу;
- відслідковувати різного роду статистику;
- формувати звітність, проводити документування навчального процесу; та ін.

Основні функції системи управління навчанням представлені на рисунку 1.



Рисунок 1 - Система управління навчанням

Прикладами систем управління навчання є Moodle, OpenedX, Blackboard і ін. Така система здатна накопичувати величезний обсяг інформації про процес навчання, містити різного роду статистику про навчання: результати тестувань, швидкість виконання завдань і багато інших даних. Аналіз такої інформації може бути застосований з метою

поліпшення освітнього процесу, розробки рекомендацій, як викладачеві, так і учневі.

У зв'язку з цим з'явилося такий напрямок як Educational Data Mining.

Educational Data Mining (EDM) - новий напрямок інтелектуального аналізу даних, засноване на зборі, аналізі та поданні даних, отриманих з освітніх параметрів, з метою оптимізації навчального процесу.

За допомогою впровадження EDM стає можливим рішення таких задач, як сегментація навчаються, прогнозування (передбачення результату навчання студента), генерація рекомендація зацікавленим особам (навчається, викладач, вищий навчальний заклад) і ін.

Система Moodle - система дистанційного навчання, орієнтована на організацію взаємодії між викладачем і студентом, при цьому система підходить як для організації дистанційних курсів, так і для організації процесу очного навчання.

Система Moodle надає можливість створювати курси, розміщуючи навчальні матеріали, а також забезпечує таку форму контролю знань як тестування.

При цьому система має спеціальними інструментами для створення тестів і проведення навчального і контрольного тестування.

Тестові завдання можуть бути різних типів: множинний вибір, питання на відповідність, з коротким відповіддю і багато інших.

Система проводить також і обробку даних тестів, вона містить в собі статистичні показники для оцінки результатів тестування, оцінки складності окремих тестових питань та ін.

Розглянемо деякі з даних показників.

- Індекс легкості  $F_i$  показує, яка частина студентів відповіла правильно на аналізовані питання.

$$F_i = \frac{\bar{x}_i}{x_i(\max)} \quad (1)$$

де  $\bar{x}_i$  - середнє значення набраних студентами балів за  $i$  - е тестове завдання,  $x_i(\max)$  - максимально можлива оцінка за  $i$  е тестове завдання.

Необхідно, щоб тест містив завдання різного рівня складності, але при цьому виключав завдання з індексом легкості рівним як 1, так і 0 (що означає, що завдання є надмірно простим або надмірно складним).

- Стандартне відхилення характеризує розкид значень оцінок, отриманих за дане завдання тесту. Стандартне відхилення дорівнює 0 означає, що всі студенти отримали за питання однакову оцінку.

- Бал випадкового вгадування - оцінка, яку міг би отримати студент при випадковому вгадуванні відповідей. Наприклад, якщо необхідно вибрати один варіант з чотирьох, то бал випадкового вгадування становитиме 25%.

- Намічена вага - вага, якій призначається тестове завдання при формуванні сценарію тесту.

$$IQW_p = 100\% \cdot \frac{x_p(\max) - x_p(\min)}{T_{\max} - T_{\min}}, \quad (2)$$

де  $x_p(\max)$ ,  $x_p(\min)$  - максимально і мінімально можливі оцінки за завдання,  $T_{\max}$ ,  $T_{\min}$  - максимальна і мінімальна можливі оцінки за тест.  $T_{\min}$  завжди рівні 0.

Параметр розраховується для «позиції» завдання в тесті, тобто на кожну позицію випадково вибирається один з декількох альтернативних питань на задану тему.

Прийняті позначення:

- $I$  - множина питань;
- $P$  - множина позицій в тесті;
- $S$  - множина спроб здачі тесту;
- $T_s$  - кількість набраних балів за тест;
- $x_p(s)$  - оцінка студента за конкретне завдання;
- $X_p(s)$  - оцінка студента за всі інші завдання.

$$T_s = \sum_{p \in P} x_p(s), \quad (3)$$

$$X_p(s) = T_s - x_p(s). \quad (4)$$

Дисперсія показника:

$$V(x_p) = \frac{1}{S-1} \sum_{s \in S} (x_p(s) - \bar{x}_p)^2. \quad (5)$$

Коваріація:

$$C(x_p, X_p) = \frac{1}{S-1} \sum_{s \in S} (x_p(s) - \bar{x}_p)(X_p(s) - \bar{X}_p)^2. \quad (6)$$

- Ефективна вага характеризує фактичну частку конкретного завдання в підсумковій оцінці студентів за тест:

$$EQW_p = 100\% \cdot \frac{\sqrt{C(x_p, T)}}{\sum_{p \in P} \sqrt{C(x_p, T)}}. \quad (7)$$

- Індекс дискримінації показує, наскільки взаємопов'язані правильність відповіді на дане питання і інші питання тесту:

$$D_p = 100\% \cdot \frac{C(x_p, X_p)}{\sqrt{V(x_p)V(X_p)}}. \quad (8)$$

Для хорошого тестового питання студенти з високими оцінками за нього також матимуть високі оцінки за тест в цілому, на що вкаже позитивне значення коефіцієнта. Якщо індекс легкості питання різниться від 50%,  $D_p$  не може бути 100%. Якщо індекс легкості близький до 0% або 100%, то  $D_p$  буде малий, а за однакової кількості індексу легкості 0% або 100%  $D_p$  буде не визначений.

- Ефективність дискримінації враховує зазначену вище особливість:

$$DE_p = 100\% \cdot \frac{C(x_p, X_p)}{C_{\max}(x_p, X_p)}, \quad (9)$$

де  $C(x_p, X_p)$  розраховується за формулою (6), для розрахунку  $C_{\max}(x_p, X_p)$  множені значень  $x_p(s)$  та  $X_p(s)$  попередньо сортуються за зростанням.

Система дистанційного навчання Moodle використовується для організації процесу навчання.

У 2018/2019 навчальному році на кафедрі «ІТС» дисципліна «Прикладні нейро-нечіткі системи» викладалася студентам 4-го курсу на російській мові.

Проміжне і підсумкове тестування проводилися за допомогою середовища дистанційного навчання Moodle. Студенти кафедри проходили три проміжних тесту.

Екзаменаційний тест проводився російською мовою.

За допомогою методів інтелектуального аналізу даних в даній роботі необхідно провести аналіз якості завдань: оцінити складність питань, виявити схожі за рівнем складності питання в тесті, визначити співвідношення результатів тестів.

У даній роботі необхідно провести рішення наступної задачі:

Необхідно провести кластеризацію тестових завдань для аналізу їх складності, виявлення схожих за рівнем складності питань в тесті. Таким чином, множення тестових завдань:

$$Q = \{q_1, q_2, \dots, q_j, \dots, q_n\}, \quad (10)$$

де  $q_j$  - тестове завдання.

Необхідно розділити досліджувану множену тестових завдань на кластери, то є побудувати дискретну множену кластерів  $C$ :

$$C = \{c_1, c_2, \dots, c_k, \dots, c_g\} \quad (11)$$

і відображення  $F$  множені  $Q$  на множену  $C$  ( $F:Q \rightarrow C$ ).

Рішення даного завдання дозволить оцінити якість тестових завдань, складність питань, а також виявити схожі за рівнем складності завдання.

У даній роботі необхідно знайти рішення задач інтелектуального аналізу даних засобами мови Python.

**Досліджуваний набір даних. Задача кластеризації тестових завдань.** Необхідно провести кластеризацію тестових завдань:

$Q = \{q_1, q_2, \dots, q_j, \dots, q_n\}$ , де  $q_j$  - тестове завдання.

Задача полягає в побудові дискретної множені кластерів  $C$ :  $C = \{c_1, c_2, \dots, c_k, \dots, c_g\}$ .

Досліджуваним набором даних в даній задачі є статистичні показники тестувань, що надаються системою Moodle, приклад яких представлений на рисунку 2.

№	Тип питання	Назва питання	Попитки	Індекс легкості	Стандартне відхилення
2	1 Случайный	Случайный (Вопросы на английском глава 1 промежу	45 64.44%		48.81%
3	2 Случайный	Случайный (Вопросы на английском глава 1 промежу	45 51.11%		50.55%
4	3 Случайный	Случайный (Вопросы на английском глава 1 промежу	45 60.00%		49.54%
5	4 Случайный	Случайный (Вопросы на английском глава 1 промежу	45 60.00%		49.54%
6	5 Случайный	Случайный (Вопросы на английском глава 1 промежу	45 48.89%		50.55%
7	6 Случайный	Случайный (Вопросы на английском глава 1 промежу	45 46.67%		50.45%
8	7 Случайный	Случайный (Вопросы на английском глава 1 промежу	45 62.22%		49.03%
9	7.1 Множественный выбор	q060 A Customer Resource Management (CRM) system	21 19.05%		40.24%
10	7.2 Множественный выбор	q063 A program whose job is to create, process and ad	16 87.50%		34.16%
11	7.3 Множественный выбор	q064 Microsoft Access includes _____ (книжк	16 68.75%		47.87%
12	7.4 Множественный выбор	q068 Which of the following database systems is not a r	14 57.14%		51.36%
13	7.5 Множественный выбор	q071 In database systems, the DBMS enforces rules at	13 23.08%		43.85%
14	7.6 Множественный выбор	q073 A database is considered self-describing because	12 83.33%		38.92%
15	7.7 Множественный выбор	q075 Which of the following statements is not true for p	18 44.44%		51.13%
16	7.8 Множественный выбор	q078 A database designed using spreadsheets from the	10 80.00%		42.16%
17	7.9 Множественный выбор	q083 A very popular development technique used by dat	18 55.56%		51.13%
18	7.1 Множественный выбор	q086 Which statement is not true about the relational m	16 25.00%		44.72%
19	7.11 Множественный выбор	q090 For database development, a challenge to the relat	17 11.76%		33.21%
20	7.12 Верно/Неверно	q002 In a database, each table stores data about a diffe	17 70.59%		46.97%
21	7.13 Верно/Неверно	q006 Data is recorded facts and figures; information is k	16 100.00%		0.00%
22	7.14 Верно/Неверно	q014 Databases based on the CODASYL standard are i	16 56.25%		51.23%
23	7.15 Верно/Неверно	q032 In an Enterprise-class database system, business	14 42.86%		51.36%
24	7.16 Верно/Неверно	q039 The description of a database's structure that is st	19 68.42%		47.76%
25	7.17 Верно/Неверно	q046 Data is associated with structured data relationships	16 80.00%		41.80%

Рисунок 2 - Статистичні показники тестування

**Аналіз даних засобами мови Python.** Дані являють собою кількісні ознаки (тип даних int, float).

Проміжне тестування №1.

Метод describe () дозволяє отримати зведену інформацію з даної таблиці (рисунок 3).

	F	StDev	RGS	IQW	EQW	D	DE
count	55.000000	55.000000	55.000000	5.500000e+01	55.000000	55.000000	55.000000
mean	58.155636	45.534909	23.181818	6.670000e+00	1.817818	23.637455	33.522182
std	20.560735	6.708393	17.245896	8.963646e-16	3.080082	25.212404	35.565313
min	11.760000	24.250000	0.000000	6.670000e+00	0.000000	-41.670000	-59.120000
25%	51.060000	42.765000	0.000000	6.670000e+00	0.000000	5.075000	6.800000
50%	61.110000	48.570000	25.000000	6.670000e+00	0.000000	21.280000	31.360000
75%	69.375000	50.160000	25.000000	6.670000e+00	4.130000	44.140000	58.515000
max	94.120000	51.450000	50.000000	6.670000e+00	8.800000	81.720000	100.000000

Рисунок 3 - Характеристика досліджуваного набору даних

Для оцінки досліджуваного числа кластерів була побудована дендрограма (рисунок 4) і залежність функції J(k) від кількості кластерів (рисунок 5). Вибір кількості кластерів відбувається на підставі найменшого зміни функції.

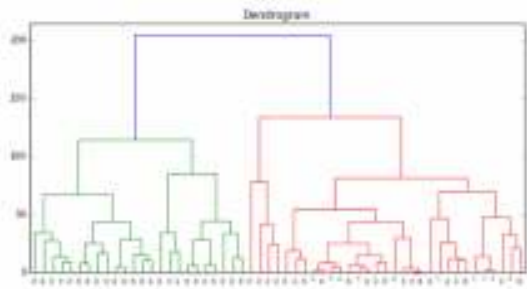


Рисунок 4 – Дендрограма

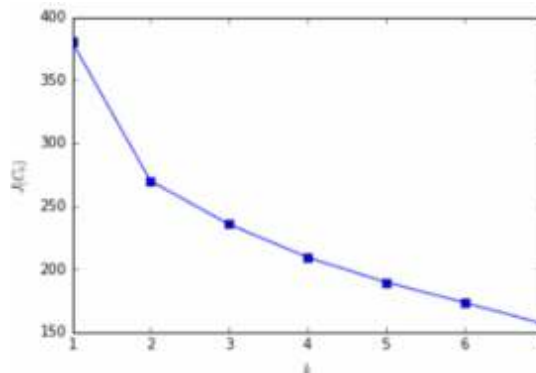


Рисунок 5 - Залежність функції  $J(k)$  від кількості кластерів

Були побудовані моделі Інтелектуального Аналізу Даних, що містять 2, 3, 4 і 5 кластерів.

Для перевірки точності моделі був застосований інструмент перехресної перевірки (рисунок 6 - 9).

```
In [25]: clf1 = svm.SVC(kernel='linear', C=1)
scores = cross_val_score(clf1, X, y, cv=2)
scores.mean()

Out[25]: 0.92791005291005291
```

Рисунок 6 - Перехресна перевірка моделі (2 кластера)

```
In [48]: clf1 = svm.SVC(kernel='linear', C=1)
scores = cross_val_score(clf1, X, y, cv=2)
scores.mean()

Out[48]: 0.94510582010582012
```

Рисунок 7 - Перехресна перевірка моделі (3 кластера)

```
In [115]: clf1 = svm.SVC(kernel='linear', C=1)
scores = cross_val_score(clf1, X, y, cv=2)
scores.mean()

Out[115]: 0.85383597883597884
```

Рисунок 8 - Перехресна перевірка моделі (4 кластера)

```
In [93]: clf1 = svm.SVC(kernel='linear', C=1)
scores = cross_val_score(clf1, X, y, cv=2)
scores.mean()

Out[93]: 0.87135278514588865
```

Рисунок 8 - Перехресна перевірка моделі (5 кластерів)

Модель ІАД, що містить 3 кластера, забезпечує найбільш точний результат.

Результати кластеризації тестового набору даних в 95% випадків збігаються з результатами кластеризації вихідних даних.

В результаті роботи алгоритму кластеризації були отримані 3 кластера.

Перший кластер містить 16 об'єктів, другий - 22 об'єкти, третій - 17 (рисунок 9).



Соотношение количества объектов  
в кластерах

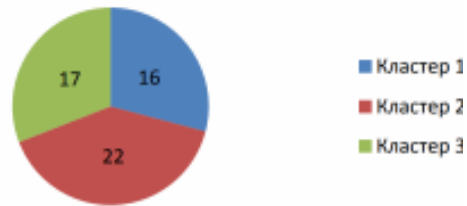


Рисунок 9 - Співвідношення кількості об'єктів в кластерах

На підставі результатів кластеризації були складені таблиці, що містять що належать певному кластеру об'єкти (рисунки 10 – 12).

	ID	F	StDev	RGS	IQW	EQW	D	DE	Cluster
0	1	57.45	49.98	0.0	6.67	6.18	3.58	4.84	1
1	2	53.19	50.44	0.0	6.67	4.09	-10.99	-14.57	1
2	3	63.83	48.57	0.0	6.67	5.87	2.35	3.15	1
3	5	51.06	50.53	0.0	6.67	5.51	-2.16	-2.90	1
4	6	55.32	50.25	0.0	6.67	4.17	-10.44	-14.02	1
5	7	61.70	49.14	0.0	6.67	6.28	5.35	6.93	1
6	8	21.05	41.89	25.0	6.67	0.00	-3.26	-3.91	1
7	10	68.75	47.87	25.0	6.67	0.00	-3.87	-5.05	1
8	16	68.75	47.87	25.0	6.67	0.00	-41.67	-59.12	1
9	24	75.00	44.72	50.0	6.67	0.00	-12.38	-16.67	1
10	27	61.11	50.16	25.0	6.67	0.00	8.22	10.64	1
11	31	55.32	50.25	0.0	6.67	6.12	2.85	3.72	1
12	38	86.36	35.13	25.0	6.67	0.00	-0.88	-1.54	1
13	49	23.81	43.64	25.0	6.67	0.00	4.80	6.67	1
14	53	92.86	26.73	50.0	6.67	0.00	-18.26	-37.70	1
15	54	86.67	35.19	50.0	6.67	0.00	-0.77	-1.35	1

Рисунок 10 - Об'єкти  
першого кластера

	ID	F	StDev	RGS	IQW	EQW	D	DE	Cluster
0	4	65.96	47.90	0.0	6.67	7.23	16.11	22.19	2
1	9	78.95	41.89	25.0	6.67	0.00	14.63	21.92	2
2	11	55.56	51.13	25.0	6.67	0.00	22.94	31.36	2
3	12	25.00	45.23	25.0	6.67	0.00	33.97	50.00	2
4	13	85.71	36.31	25.0	6.67	0.00	25.07	33.33	2
5	14	52.94	51.45	25.0	6.67	0.00	21.12	29.63	2
6	15	85.71	36.31	25.0	6.67	0.00	25.19	47.50	2
7	18	18.75	40.31	25.0	6.67	0.00	9.29	14.09	2
8	20	94.12	24.25	50.0	6.67	0.00	13.30	40.35	2
9	25	41.18	50.73	50.0	6.67	0.00	16.25	22.73	2
10	26	81.25	40.31	25.0	6.67	0.00	26.46	40.74	2
11	28	63.83	48.57	0.0	6.67	8.80	34.78	47.32	2
12	29	59.57	49.61	0.0	6.67	7.48	16.71	21.54	2
13	30	55.32	50.25	0.0	6.67	8.64	29.88	39.67	2
14	32	55.32	50.25	0.0	6.67	6.99	10.62	14.21	2
15	33	63.83	48.57	0.0	6.67	7.18	14.77	19.29	2
16	34	63.83	48.57	0.0	6.67	7.94	23.44	31.76	2
17	35	51.06	50.53	0.0	6.67	7.50	15.67	20.79	2
18	41	92.31	27.17	25.0	6.67	0.00	29.44	58.06	2
19	42	28.00	45.83	25.0	6.67	0.00	11.48	15.73	2
20	48	31.25	47.87	25.0	6.67	0.00	25.57	35.14	2
21	50	52.63	51.30	25.0	6.67	0.00	21.28	29.89	2

Рисунок 11 - Об'єкти  
другого кластера

	ID	F	StDev	RGS	IQW	EQW	D	DE	Cluster
0	17	33.33	48.80	25.0	6.67	0.0	56.85	81.25	3
1	19	62.50	50.00	50.0	6.67	0.0	44.78	58.97	3
2	21	55.00	51.04	50.0	6.67	0.0	43.50	55.56	3
3	22	37.50	50.00	50.0	6.67	0.0	58.30	87.10	3
4	23	64.71	49.26	50.0	6.67	0.0	57.61	74.05	3
5	36	60.00	50.71	25.0	6.67	0.0	67.59	87.95	3
6	37	78.95	41.89	25.0	6.67	0.0	54.29	78.57	3
7	39	81.82	39.48	25.0	6.67	0.0	45.19	72.27	3
8	40	64.71	49.26	25.0	6.67	0.0	68.73	100.00	3
9	43	11.76	33.21	25.0	6.67	0.0	39.65	55.84	3
10	44	70.00	47.02	25.0	6.67	0.0	52.97	73.51	3
11	45	33.33	49.24	25.0	6.67	0.0	81.72	100.00	3
12	46	64.71	49.26	25.0	6.67	0.0	42.20	56.78	3
13	47	61.11	50.16	25.0	6.67	0.0	45.30	59.40	3
14	51	28.57	46.29	25.0	6.67	0.0	56.65	76.85	3
15	52	75.00	44.23	50.0	6.67	0.0	45.99	69.23	3
16	55	31.25	47.87	50.0	6.67	0.0	58.10	90.12	3

Рисунок 12 - Об'єкти  
третього кластера

Для оцінки основних характеристик кластерів був застосований метод describe() (рисунки 13-15).

	F	StDev	RGS	IQW	EQW	D	DE
count	16.000000	16.000000	16.000000	1.600000e+01	16.000000	16.000000	16.000000
mean	61.389375	45.147500	18.750000	6.670000e+00	2.388750	-4.845625	-7.555000
std	19.757486	7.079462	19.354917	9.173067e-16	2.860317	12.259698	18.250568
min	21.050000	26.730000	0.000000	6.670000e+00	0.000000	-41.670000	-59.120000
25%	54.787500	43.202500	0.000000	6.670000e+00	0.000000	-10.577500	-14.157500
50%	61.405000	48.220000	25.000000	6.670000e+00	0.000000	-1.520000	-2.220000
75%	70.312500	50.182500	25.000000	6.670000e+00	5.600000	3.032500	4.000000
max	92.860000	50.530000	50.000000	6.670000e+00	6.280000	8.220000	10.640000

Рисунок 13 - Характеристики  
першого кластера

	F	StDev	RGS	IQW	EQW	D	DE
count	22.000000	22.000000	22.000000	2.200000e+01	22.000000	22.000000	22.000000
mean	59.185455	44.742727	16.181818	6.670000e+00	2.807273	20.825455	31.233636
std	21.585732	7.764374	15.777127	9.090796e-16	3.821255	7.449512	12.488286
min	18.750000	24.250000	0.000000	6.670000e+00	0.000000	9.290000	14.090000
25%	51.482500	40.705000	0.000000	6.670000e+00	0.000000	14.995000	21.635000
50%	57.565000	48.235000	25.000000	6.670000e+00	0.000000	21.200000	30.625000
75%	75.702500	50.250000	25.000000	6.670000e+00	7.217500	25.475000	40.180000
max	94.120000	51.450000	50.000000	6.670000e+00	8.800000	34.780000	58.060000

Рисунок 14- Характеристики  
другого кластера

	F	StDev	RGS	IQW	EQW	D	DE
count	17.000000	17.000000	17.000000	1.700000e+01	17.0	17.000000	17.000000
mean	53.779412	46.924706	33.823529	6.670000e+00	0.0	54.084118	75.144706
std	20.420203	4.765214	12.314805	9.155134e-16	0.0	11.159361	14.724161
min	11.760000	33.210000	25.000000	6.670000e+00	0.0	39.650000	55.560000
25%	33.330000	46.290000	25.000000	6.670000e+00	0.0	45.190000	59.400000
50%	61.110000	49.240000	25.000000	6.670000e+00	0.0	54.290000	74.050000
75%	64.710000	50.000000	50.000000	6.670000e+00	0.0	58.100000	87.100000
max	81.820000	51.040000	50.000000	6.670000e+00	0.0	81.720000	100.000000

Рисунок 15 - Характеристики третього кластера

Розглянемо основні характеристики кластерів. Значення індексу легкості кластерів:

$$F_1 \in [21.05, 92.86],$$

$$F_2 \in [18.75, 94.12],$$

$$F_3 \in [11.76, 81.82].$$

Значення стандартного відхилення кластерів:

$$StDev_1 \in [26.73, 50.53],$$

$$StDev_2 \in [24.25, 51.45],$$

$$StDev_3 \in [33.21, 51.04].$$

Значення індексу дискримінації кластерів:

$$D_{p1} \in [-41.67, 8.22],$$

$$D_{p2} \in [9.29, 34.78],$$

$$D_{p3} \in [39.65, 81.72].$$

Значення індексу легкості і значення стандартного відхилення кластерів мають схожі значення. Основна відмінність кластерів укладає в значенні індексу дискримінації. Перший кластер включає в себе об'єкти з від'ємним значенням індексу дискримінації, що вказує на відсутність взаємозв'язку правильності відповіді на дані тестові питання з іншими відповідями на питання тестування. Даний кластер містить 16 тестових завдань, наприклад, питання №1, що відноситься до типу «випадковий», і питання №7.17, що відноситься до типу «вірно / невірно». другий кластер містить об'єкти з невеликим значенням індексу дискримінації  $Dp \in [9.29, 34.78]$  В літературі відзначається, якщо індекс легкості близький до 0% або 100%, то індекс дискримінації буде малий. Третій кластер містить об'єкти з середнім значенням індексу дискримінації.

Таке дослідження було проведено і для проміжного тестування №2 та екзаменаційного тестування.

Рішення завдання засобами Python дозволило проаналізувати тестові завдання, характеризуються відсутністю взаємозв'язку правильності відповіді з іншими відповідями на питання тестування.

**Висновки.** В ході написання даної роботи було проведено дослідження методів інтелектуального аналізу даних засобами мови Python.

В роботі були вивчені завдання інтелектуального аналізу даних, методи їх вирішення, а також кошти мови Python для вирішення цих завдань.

В якості досліджуваного набору даних були обрані результати тестування студентів кафедри «ІТС» за 2018/2019 рік по курсу «Прикладні нейро-нечіткі системи». Студенти проходили 2 проміжних тесту і один екзаменаційний тест в системі дистанційного навчання Moodle.

У даній роботі була поставлена і вирішена задача кластеризації тестових завдань. Дана задача була вирішена засобами мови Python, проте результати кластеризації не співпали. Рішення завдання засобами Python дозволило проаналізувати тестові завдання, які характеризуються відсутністю взаємозв'язку правильності відповіді з іншими відповідями на питання тестування, а також виявити співвідношення таких тестових завдань для тестувань, що проводяться.

#### **ЛІТЕРАТУРА / ЛИТЕРАТУРА**

1. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. Учебное пособие. / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. - М.:БХВ-Петербург, 2007. – 331 с.
2. Нестеров С.А. Базы данных. Интеллектуальный анализ данных. Учебное пособие. / С.А. Нестеров. - СПб.: Изд-во Политехн. ун-та, 2011. –272 с.
3. Чубукова И.А. Data Mining. Задачи Data Mining. Классификация и кластеризация. Режим доступа:  
<http://www.intuit.ru/studies/courses/6/6/lecture/166?page=4>.
4. Структуры интеллектуального анализа данных (службы Analysis Services - интеллектуальный анализ данных). Режим доступа:  
<https://msdn.microsoft.com/ru-ru/library/ms174757>.
5. Путівник мовою програмування Python. Режим доступа:  
<https://pythonguide.rozh2sch.org.ua/>

#### **REFERENCES**

1. Barsegyan A.A. Tehnologii analiza dannyih: Data Mining, Visual Mining, Text Mining, OLAP. Uchebnoe posobie. / A.A. Barsegyan, M.S. Kupriyanov, V.V. Stepanenko, I.I. Holod. - M.:BHV-Peterburg, 2007. – 331 s.

2. Nesterov S.A. Bazyi dannyih. Intellektualnyiyy analiz dannyih. Uchebnoe posobie. / S.A. Nesterov. - SPB.: Izd-vo Politehn. un-ta, 2011. –272 s.
3. Chubukova I.A. Data Mining. Zadachi Data Mining. Klassifikatsiya i klasterizatsiya. Rezhim dostupa:  
<http://www.intuit.ru/studies/courses/6/6/lecture/166?page=4>.
4. Strukturyi intellektualnogo analiza dannyih (sluzhbyi Analysis Services - intellektualnyiyy analiz dannyih). Rezhim dostupa:  
<https://msdn.microsoft.com/ru-ru/library/ms174757>.
5. Putivnik movoyu programuvannya Python. Rezhim dostupa:  
<https://pythonguide.rozh2sch.org.ua/>

Received 12.03.2020.

Accepted 16.03.2020.

**Исследование методов интеллектуального анализа данных  
для обработки результатов тестирования**

*Работа посвящена исследованию методов интеллектуального анализа данных для обработки результатов тестирования с использованием языка Python.*

*Целью данной работы является исследование методов интеллектуального анализа данных средствами языка Python.*

*Для достижения поставленной цели необходимо решение следующих задач:*

- исследование этапов интеллектуального анализа данных, особенностей подготовки и изучения данных;
- исследование задач интеллектуального анализа данных и методов их решения;
- исследование средств языка Python для решения задач интеллектуального анализа данных;
- практическое решение задач ИАД: построение моделей интеллектуального анализа данных средствами языка Python;
- анализ качества построенных моделей.

*В данной работе был проведен интеллектуальный анализ данных, полученных из системы управления обучением Moodle.*

**Research methods of data mining for processing test results**

*The work is devoted to the study of data mining methods for processing test results using the Python language.*

*The aim of this work is to study methods of data mining using Python.*

*To achieve this goal it is necessary to solve the following tasks:*

- study of the stages of data mining, features of the preparation and study of data;
- study of data mining problems and methods for solving them;
- research of Python language tools for solving data mining problems;
- a practical solution to the problems of IAD: building data mining models using Python;
- analysis of the quality of the constructed models.

*In this work, an intellectual analysis of the data obtained from the Moodle learning management system was carried out.*

*In the work, data mining problems, methods for solving them, and also Python tools for solving these problems were studied.*

*As the studied data set, the results of testing students of the ITS department for 2018/2019 at the course "Applied Neuro-Fuzzy Systems" were selected. Students took 2 intermediate tests and one exam test in the Moodle distance learning system.*

*In this paper, the task of clustering test tasks was posed and solved. This task was solved using the Python language, but the clustering results did not match. Solving the problem using Python tools allowed us to analyze test tasks, which are characterized by the absence of a correlation between the correctness of the answer and other answers to testing questions, and also to identify the relationship of such test tasks for testing conducted.*

**Островская Екатерина Юрьевна** - к.т.н., доцент кафедри інформаційних технологій та систем, Національна металургійна академія України.

**Островська Катерина Юріївна** – к.т.н., доцент кафедри Інформаційних технологій та систем, Національна металургійна академія України.

**Ostrovska Kateryna** - candidate of technical sciences, associate professor of the Department of Information Technology and System, National Metallurgical Academy of Ukraine.