УДК 378.5:004.67
DOI https://doi.org/10.32782/cusu-pmtp-2024-2-12

ДО ПИТАННЯ ВИБОРУ ВІЛЬНОПОШИРЮВАНИХ ЗАСОБІВ ДЛЯ ВИВЧЕННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ У ЗАКЛАДАХ ВИЩОЇ ОСВІТИ

Лупан Ірина Володимирівна,
кандидат педагогічних наук, доцент,
доцент кафедри інформатики, програмування, штучного інтелекту та технологічної освіти
Центральноукраїнського державного університету
імені Володимира Винниченка
ORCID ID: 0000-0002-4791-0445

Підгорна Тетяна Володимирівна,
доктор педагогічних наук, доцент,
професор кафедри комп'ютерних наук та інформаційних систем
Державного торговельно-економічного університету
ORCID ID: 0000-0002-1414-3489

Інтелектуальний аналіз даних (ІАД) є одним з найважливіших напрямів у розвитку інформаційних технологій, тому дисципліни, пов'язані з ІАД, включено в освітній стандарт підготовки фахівців у галузі комп'ютерних наук. Однак вибір програмних засобів для навчання залишається актуальним, оскільки, з одного боку, засоби, які зазвичай використовують у практичній діяльності підприємств, великих ІТ компаній, агенцій, що спеціалізуються на аналізі даних, є пропрієтарними і досить дорогими, а з іншого боку – у майбутніх фахівців повинні бути сформовані знання і навички щодо застосування основних методів та алгоритмів аналізу даних, особливостей підготовки даних до того чи іншого виду аналізу, форматів представлення результатів аналізу та уміння інтерпретувати отримані результати. У такому разі для навчальних цілей цілком прийнятним буде використання безкоштовних засобів за умови відповідності їхніх функціональних можливостей навчальним цілям дисципліни. У статті досліджуються види програмного забезпечення – табличні процесори, спеціалізовані пакети та мови програмування – на предмет придатності до використання під час навчання аналізу даних. У статті наведено порівняння функціонала деяких з вказаних засобів; наведено приклади їх використання під час аналізу, зокрема кластерного, за допомогою RapidMiner, KNIME, Orange, JASP, R. Зроблено висновок про можливість використання вільнопоширюваного програмного забезпечення за умови відповідності його функціонала цілям освітнього процесу та наведено результати педагогічного експерименту, в якому було доведено, що якість засвоєння навчального матеріалу не залежить від того, який програмний засіб застосовано в процесі вивчення дисципліни. Однак, добираючи програмні засоби, доцільно враховувати їхню вартість та функціонал (охоплення методів аналізу, засоби візуалізації, якість отримуваних результатів тощо).

Ключові слова: *інтелектуальний аналіз даних, пакети аналізу даних, вільнопоширюване програмне забезпечення, кластерний аналіз, підготовка фахівців.*

Lupan Iryna, Pidhorna Tetiana. To the Question of Choosing Free Tools to Study Data Mining Courses in Higher Education Institutions

Data Mining (DM) is one of the most important areas in the development of information technologies, so disciplines related to DM are included into the educational standard for the specialists in the field of computer sciences. However, the choice of training software are relevant, because, on the one hand, tools that are commonly used in the practical activity of enterprises, large IT companies, agencies specializing on data analysis are proprietary and quite expensive. On the other hand, the future specialists should be formed knowledge and skills in the use of basic methods and algorithms of data analysis, features of data preparation for the different types of analysis, formats for presenting the results of the analysis and the ability to interpret the results. In this case, the usage of free means will be quite acceptable for the educational purposes, provided that their functionality complianes with the objectives of the discipline. The article examines such types of software as spreadsheet programs, specialized packages and

programming languages - for the usage of data analysis during the training. At the article some of these tools were compared. Examples of using SPSS, RapidMiner, Knime, Orange, Jasp and R for cluster analysis were given. However, the results of the pedagogical experiment show that the quality of learning of educational material does not depend on which software were used during studying the discipline. Therefore, when choosing software, it is advisable to evaluate their cost and functionality (coverage of methods of data mining, visualization tools, quality of results, etc.). A conclusion about the possibility of using free software if its functionality matchs the objectives of the learning was made.

Key words: *data mining; analytical packages; free software; cluster analysis, training of specialists.*

Вступ. Інтелектуальний аналіз даних як сучасна концепція аналізу різноманітних інформаційних матеріалів з використанням як можливостей людського інтелекту, так і засобів сучасних інформаційних технологій та реалізованих з їх використанням методів пошуку латентних закономірностей, прихованих у наявних даних, набуває все більшого розвитку і значення, особливо із збільшеннями обсягів накопичуваних даних. Відповідно зростає потреба у фахівцях, здатних здійснювати кваліфікований аналіз даних з використанням цих методів, розробляти програмне забезпечення для автоматизації цього процесу, удосконалювати існуючі аналітичні процедури та розробляти нові методи та підходи до аналізу даних.

У зв'язку з цим у редакції стандарту вищої освіти, прийнятого у 2019 році [1], «здатність до інтелектуального аналізу даних» визначено як одну з ключових фахових компетентностей для бакалаврів спеціальності 122 «Комп'ютерні науки».

Методи інтелектуального аналізу даних широко використовують у найрізноманітніших галузях людської діяльності: економіці, банківській справі, телекомунікаціях, автомобілебудуванні, авіаперевезеннях, молекулярній генетиці та генній інженерії, прикладній хімії, біоінформатиці, медицині, освіті [2; 3] та інших.

У згаданому вище стандарті здатність до інтелектуального аналізу даних передбачає «знання методів аналітичної обробки … для задач класифікації, прогнозування, кластерного аналізу, пошуку асоціативних правил», а також «використання програмних інструментів підтримки аналізу даних», «використання технологій DataMining, TextMining, WebMining для інтелектуального аналізу даних» [1]. В оглядах [4–11] найбільш часто використовуваними засобами інтелектуального аналізу даних (Data Mining) називають RapidMiner, IBM SPSS Modeler, Weka, KNIME, Orange, Apache Mahout, SAS Enterprise Mining. Проблема полягає у тому, щоб обрати для навчання засоби, функціонал яких забезпечує опанування визначених стандартом знань та умінь, але таких, що не вимагають складної та матеріальновитратної процедури ліцензування (як слушно зазначають О.А. Журан та К.В. Донченко [12], більшість з перелічених засобів є досить дорогими, а отже недоступними для закладів освіти). Ще одним міркуванням є можливості використання обраних засобів для навчання як в очному, так і у дистанційному або змішаному форматі: при цьому доведеться відмовитися від засобів, використання яких не передбачає індивідуальних дозволів для студентів навіть тих навчальних закладів, які мають ліцензійну угоду з правовласниками даного програмного забезпечення, та враховувати можливості студентів на встановлення і використання необхідного програмного забезпечення у домашніх умовах.

Недостатньо досліджені також питання пов'язані із вибором типу програмного забезпечення (табличні процесори, пакети спеціального призначення, мови програмування), яке доцільно використовувати в освітньому процесі під час вивчення інтелектуального аналізу даних.

Метою даної роботи є дослідження функціоналу вільнопоширюваного програмного забезпечення інтелектуального аналізу даних з погляду його відповідності цілям освітнього процесу та підтвердження гіпотези про те, що вибір засобів аналізу даних не впливає на рівень опанування навчального матеріалу (на прикладі застосування процедур кластерного аналізу за допомогою пакета KNIME та мови R).

Аналіз досліджень і публікацій. Під час виконання дослідження було здійснено аналіз наукових публікацій і навчальних програм з відповідних курсів, виконано порівняльний аналіз функціоналу програмного забезпечення з аналізу даних, проведено педагогічний експеримент.

Також було проведено анкетування серед фахівців, які вивчали аналіз даних та так чи інакше застосовують методи аналізу даних у професійній діяльності. На жаль, вибірка опитаних виявилася нерепрезентативною (до опитування долучилися тільки 17 осіб), оскільки таких фахівців небагато і вони розпорошені за різними компаніями та установами. Однак проведене опитування можна розглядати як пілотне для вивчення поставленої проблеми. Можна було розширити вибірку за рахунок долучення викладачів ЗВО, проте вони не становлять цільову аудиторію для нашого дослідження.

Отже, в анкетуванні взяли участь працівники IT-компаній (47% опитаних), а також економісти, фінансисти, співробітники управління статистики тощо з різним досвідом роботи та стажем (від одного до 20 років). Частина з них вивчала аналіз даних в університеті. Серед засобів аналізу даних, які найчастіше доводиться використовувати у професійній діяльності, назвали MS Excel (88%), SPSS (29%), Python (35%), R (29%) – можна було вказати декілька засобів. Окремі опитані вказали такі засоби як Google Spreadsheets, 1C, Statistica, KNIME, JASP, MatLab, SQL.

Аналіз навчальних програм і силабусів з курсу інтелектуального аналізу даних показав, що при викладанні цієї дисципліни у закладах вищої освіти існують різні варіанти використання програмного забезпечення. Так у Запорізькому національному університеті (Україна) та Università di Bologna (Італія) використовують табличні процесори, у Національному технічному університеті України «Київський політехнічний інститут імені Ігоря Сікорського» (Україна), Західноукраїнському національному університеті (Україна) орієнтуються на прикладні пакети, у Чернівецькому національному університеті імені Юрія Федьковича (Україна) та у курсі на платформі Prometheus – на мови програмування; а у Державному вищому навчальному закладу «Ужгородський національний університет» (Україна) та курсах на платформі Udemy – використовують їх комбінації.

У посібниках [13–14] наведено наш досвід використання у навчанні різноманітних програмних засобів аналізу даних. У посібнику [13] розглянуто базовий функціонал та особливості застосування пакетів MS Excel, SPSS та Statistica. Перевагами пакета MS Excel є його поширеність та обізнаність більшості потенційних користувачів з основами роботи. Пакет і досі залишається одним з основних інструментів економічного аналізу [15], що підтверджується і нашим опитуванням, однак його базовий функціонал не забезпечує виконання таких важливих процедур інтелектуального аналізу даних як, наприклад, кластерний, факторний та дискримінантний аналіз. Для вивчення цих методів необхідно додатково встановлювати розширення, що потребує витрат на ліцензування. Однак, знати межі застосування MS Excel та максимально його використовувати принаймні для статистичного аналізу буде корисно для тих, хто працює з даними. Такі ж міркування можна навести і стосовно Google Spreadsheets та Libre Office Calc: це пакети безкоштовні, але їхній функціонал на здійснення більшості процедур інтелектуального аналізу даних не налаштований.

Пакети SPSS та Statistica це статистичні пакети, за інтерфейсом трохи схожі на розглянуті вище засоби опрацювання електронних таблиць, що полегшує деяким користувачам опанування особливостей роботи з ними, крім того IBM SPSS Modeler називають як один з найпопулярніших засобів аналізу даних. Однак для використання цих пакетів потрібно отримати ліцензію, що може бути занадто дорогим задоволенням для навчального закладу (а от установи, в яких доведеться працювати майбутнім фахівцям, такі ліцензії можуть мати). Академічну ліцензію на SPSS слід підтверджувати щороку, а персональна студентська ліцензія (Campus License) на теренах нашої країни не діє. Для ознайомлення з базовим функціоналом IBM SPSS Inc. можна скористатися 30-денною версією, проте це не дуже зручно в умовах освітнього процесу.

У посібнику [14] запропоновано використовувати у навчанні пакети RapidMiner [16] або KNIME [17]. Це спеціалізовані пакети аналізу даних, що мають схожий інтерфейс, який істотно відрізняється від інтерфейсів вище згаданих програмних засобів. Весь процес аналізу, починаючи з процедур очищення та перетворення початкових даних, формується з окремих вузлів, які можна налаштувати за потребами користувача. Однак, починаючи з деякого часу, особливо коли після придбання RapidMiner компанією Altair (Nasdaq: ALTR) у березні 2023 року пакети аналітики даних і штучного інтелекту були об'єднані в одну платформу, для користуванням пакетом Altair RapidMiner потрібно отримати ліцензію. Втім умови для студентів досить вигідні: для використання у некомерційних цілях можна отримати річну безкоштовну ліцензію з можливістю поновлення. Пакет KNIME поки залишається безкоштовним, що робить його більш привабливим. Однак в процесі викладання з'ясувалося, що незважаючи на велику кількість прикладів, які надаються користувачу, пакет недостатньо документований.

На наш погляд привабливими для освітнього процесу є менш популярні поки що але безкоштовні пакети Orange Data Mining [18] та Jasp [19].

Аналітична система Orange Data Mining – це програма з відкритим вихідним кодом для машинного навчання, інтелектуального аналізу та візуалізації даних з великим набором функцій, яка розробляється Лабораторією біоінформатики Люблянського університету. У програмному забезпеченні Orange Data Mining застосовується візуальне програмування, в рамках якого аналітичні процедури створюються шляхом зв'язування вбудованих або розроблених користувачем блоків (віджетів), що дуже нагадує створення процесу виконання у Rapid Miner або у KNIME. Крім того, розробники зазначають, що досвідчені користувачі можуть використовувати Orange Data Mining як програмну бібліотеку Python для маніпулювання даними та створення нових віджетів.

JASP (розшифровується як Jeffreys's Amazing Statistics Program на знак визнання піонера байєсівських висновків сера Гарольда Джеффріса) – це безкоштовна програма з відкритим кодом для статистичного аналізу, яка підтримується Амстердамським університетом. Пакет забезпечує простий інтерфейс, інтуїтивно зрозумілий аналіз з обчисленнями в реальному часі та відображення всіх результатів у форматі, знайомому користувачам SPSS. Він пропонує стандартні процедури аналізу як у класичній, так і у байєсівській формі. Перелік аналітичних процедур, реалізованих на даний час, наведено на сторінці https://jasp-stats.org/features/. JASP читає зокрема формати .csv, .txt, .ods, .dta, .sav, .por, .jasp, а таблиці можна експортувати з JASP у формат LaTeX. Пакет супроводжується наборами даних і методичними посібниками, його вже використовують викладачі 292 університетів з 67 різних країн.

Безумовно, для майбутніх професійних програмістів можливо більш звичними засобами будуть мови програмування Python або R: обидві мови мають велику кількість бібліотек, завдяки чому охоплюють надзвичайно широкий спектр методів аналізу даних; обидві є безкоштовними у використанні; обидві мають розвинені засоби редагування та дозволяють виконувати аналіз у режимі інтерпретації, тобто процес опрацювання даних можна здійснювати покроково і переглядати результати виконання окремих кроків. Тим не менше, на нашу думку, при вивченні дисципліни «Інтелектуальний аналіз даних» важливо звертати увагу студентів саме на її аналітичний аспект: на вимоги до подання вхідних даних, на особливості представлення результатів різних методів аналізу, на інтерпретацію отриманих результатів тощо.

Результати. Розглянемо особливості та результати виконання кластерного аналізу з використанням різного інструментарію на невеликому наборі даних, отриманих у психологічному тестуванні (табл. 1). Тут значення від 0 до 4 балів означають низький, а 5–8 балів – помірний рівень вираженості якості, тобто в цілому адаптивну поведінку. Бали 9–12 – високий рівень, а бали 13–16 – поведінку екстремальну до патології.

Таблиця 1

Приклад масиву даних для кластерного аналізу

код	лідерство	впевненість	вимогливість	скептицизм	поступливість	довірливість	добросердя	чуйність
1	2	2	3	3	10	10	8	7
2	15	9	8	8	8	7	13	11
3	9	5	9	9	8	3	9	4
4	5	5	9	9	3	9	8	4
5	8	6	10	10	7	7	9	7
6	1	4	2	2	9	7	8	9
7	7	4	5	15	15	14	12	12
8	10	10	10	10	15	13	14	14
9	11	7	9	9	7	6	8	6
10	11	4	9	9	9	8	9	10
11	10	8	8	8	9	11	11	12
12	6	2	8	8	8	9	12	9
13	11	9	6	6	5	5	10	4
14	4	5	4	4	9	10	10	9

Виконання кластерного аналізу з використанням пакета SPSS детально розглянуто нами у посібнику [13]. Процедура виконання кластерного аналізу у пакетах Rapid Miner та KNIME розглянуто у посібнику [14]. Також у цитованих посібниках розглядаються особливості виконання за допомогою вказаних програмних засобів дискримінантного та факторного аналізу, процедур описової статистики тощо. У посібнику [14] крім того розглянуто порядок здійснення аналізу асоціативних правил засобами пакетів Rapid Miner та KNIME. Тож розглянемо процедури виконання кластерного аналізу за допомогою пакетів JASP, OrangeDataMining та мови R.

Кластерний аналіз в JASP.

Для виконання у пакеті JASP збережемо дані табл. 1 у форматі .ods або .csv. Інколи типи змінних (числовий, порядковий, номінативний) розпізнаються неправильно, але це легко виправити після завантаження у JASP. Далі потрібно лише вибрати бажаний вид аналізу з верхнього меню та виконати налаштування. Якщо у базовому меню потрібний вид аналізу відсутній, слід пошукати його у меню додаткових модулів.

А Б В

Рис. 1. Налаштування та результати виконання кластерного аналізу (JASP)

На рисунку 1 показано, що на екран можна виводити почергово або одночасно у різних комбінаціях таблицю з даними (А), панель налаштування параметрів методу (Б) або звіт з результатами аналізу (В).

Кожен елемент звіту, або увесь звіт, можна скопіювати та вставити у текстовий документ, причому текстові елементи (заголовки блоків, таблиці) можна буде форматувати. Для графічних об'єктів є можливість змінювати деякі параметри у JASP.

Кластерний аналіз в Orange Data Mining.

Для виконання кластерного аналізу в Orange вихідні дані були збережені у форматі CSV (без заголовків стовпців). Далі, за допомогою вбудованих віджетів було створено схему виконання обчислень, як показано на рис. 2.

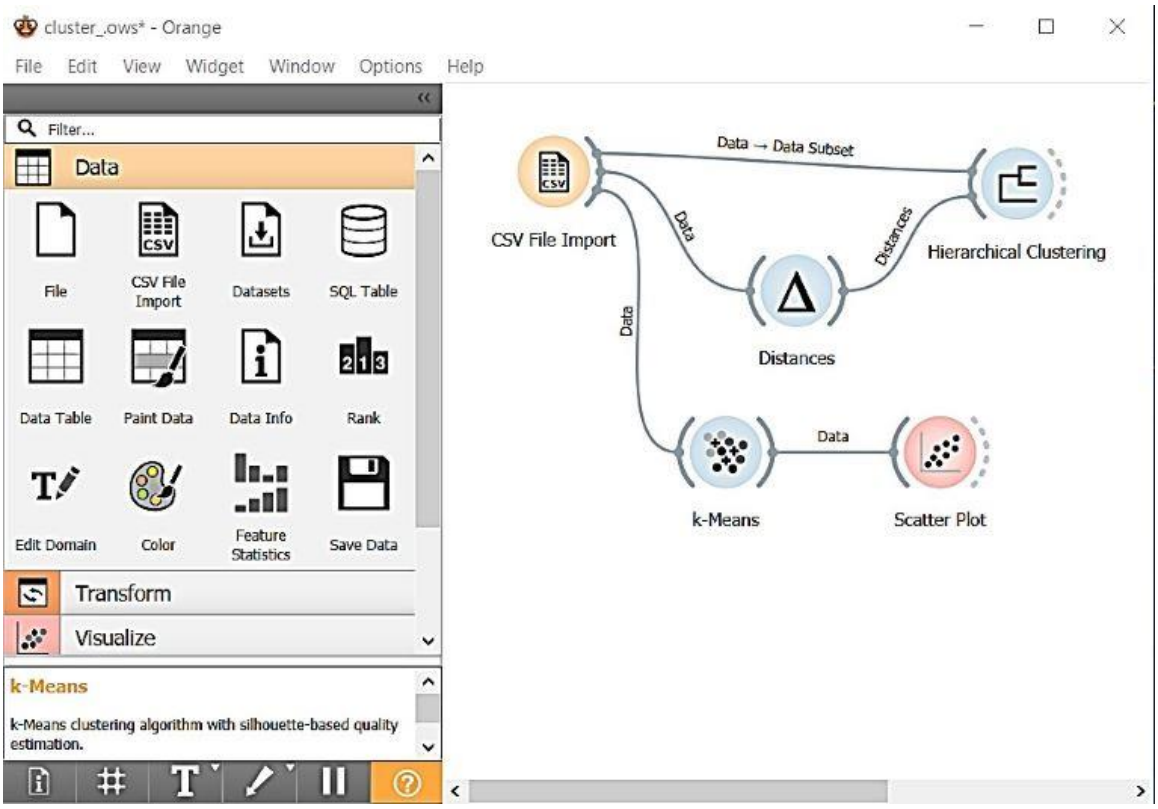

Рис. 2. Налаштування потоку виконання кластерного аналізу (Orange)

Усі віджети налаштовуються за потребами користувача: для даних можна встановити тип кодування, вид розділового знаку та тип даних (рис. 3, 4); для ієрархічного кластерного аналізу (попередньо визначивши метрику обчислення відстаней та з'ясувавши, що буде порівнюватися – рядки чи стовпці) налаштовують спосіб об'єднання кластерів (на рис. 6 – це метод Варда); для аналізу k-means налаштовують кількість кластерів (рис. 5), приналежність до яких буде відображено на діаграмі розсіювання у координатах двох вибраних змінних (рис. 7). Результати кластерного аналізу: дендрограму, діаграму розсіювання та оцінки силуету для різної кількості кластерів, – можна вивести у звіт та зберегти у форматах html, pdf, або у власному форматі пакета Orange.

Як видно з рисунків, засоби візуалізації в Orange досить наочні – наприклад, кластери на дендрограмі виділяються кольором, крім того їхню кількість можна варіювати, переміщуючи

вертикальну пунктирну лінію; процес побудови схеми виконання аналізу не складний, якщо орієнтуєшся в методах. Звичайно, щоб максимально використовувати інструментарій пакета, потрібно з ним більше попрацювати, але навіть з першого погляду враження позитивне. До того ж на сайті представлено велику кількість навчального відео, є документація, блог, приклади застосування тощо.

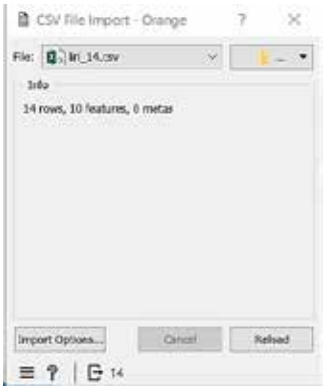

Рис. 3. Налаштування віджету CSV File Import

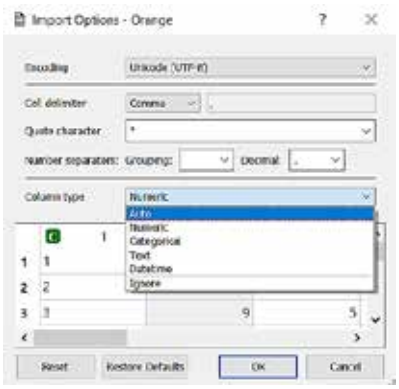

Рис. 4. Налаштування типу змінної

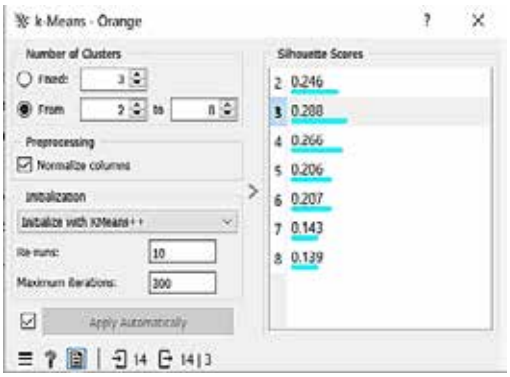

Рис. 5. Визначення кількості кластерів за методом силуету

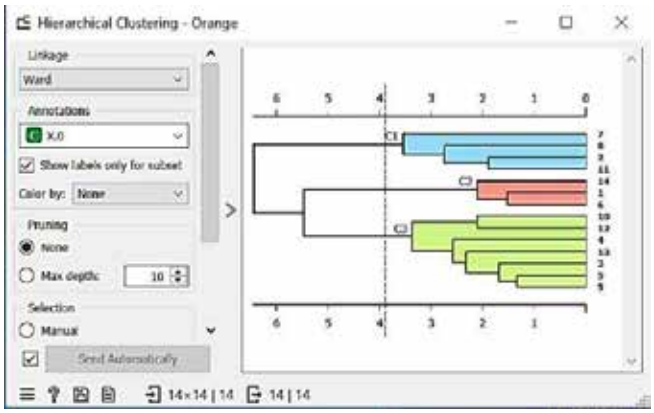

Рис. 6. Дендрограма

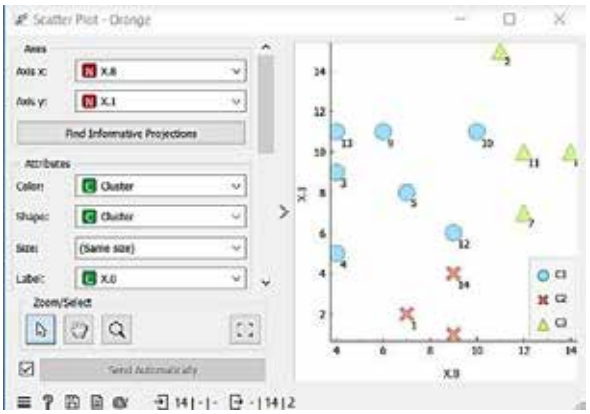

Рис. 7. Діаграма розсіювання

Виконання кластерного аналізу засобами R

Для виконання аналізу засобами мови R у середовищі RStudio спочатку необхідно створити фрейм даних. Щоб скористатися даними, які зберігаються у форматі CSV, слід виконати функцію read.csv(). Щоб не прописувати щоразу шлях до файлу, краще на початку роботи налаштувати робочу папку за допомогою функції setwd(). Нехай, наприклад, файл з даними розміщено у папці d:\R_examples:

```
> setwd('d:\\R_examples') # налаштування робочої папки
> dir() # перегляд вмісту поточної папки
> liri<-read.csv(file="liri_14.csv", header=TRUE, sep=";", dec=".", row.
names="id") # створення фрейма даних liri з файла з використанням існуючих
заголовків стовпців для іменування змінних, а значень стовця «id» для
іменування рядків, тобто об’єктів. Як розділовий знак вказано кому.
```

Далі можна переконатись у створенні фрейма, переглянувши перші рядки:

```
> head(liri)
    var1 var2 var3 var4 var5 var6 var7 var8
s_1  2    2    3    3    10   10   8    7
s_2  15   9    8    8    8    7    13   11
s_3  9    5    9    9    8    3    9    4
s_4  5    5    9    9    3    9    8    4
s_5  8    6    10   10   7    7    9    7
s_6  1    4    2    2    9    7    8    9
```

В мові R для кластеризації методом k-середніх застосовують функцію kmeans з пакета stats:

```
kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong",
"Lloyd", "Forgy", "MacQueen"), trace=FALSE)
```

де, згідно до формату використання,

x – матриця даних (фрейм), кожен рядок якої є вектором ознак чергового об'єкта спостереження;

centers – кількість кластерів k або набір початкових центрів кластерів. Якщо набір початкових центрів не задано, то обирають k випадкових об'єктів спостереження.

Решта параметрів необов'язкова, детальний опис функції наводиться у довідці.

Доступ до сформованих значень можна отримати шляхом запису їхньої назви в подвійних квадратних дужках справа від змінної, яка містить модель, або через знак долара. Наприклад, «a[["cluster"]]» або «a$cluster». Оптимальна кількість кластерів можна визначити так:

```
> cl <- 0 # ініціалізація першого значення вектора cl
> for (i in 1:13) cl[i] <- sum(kmeans(liri, centers=i)$withinss) # обчислення
суми внутрішньокластерних сум квадратів (kmeans$withinss) для різної кількості
кластерів – визначення ординат точок майбутнього графіка
> plot(1:13, cl, type="b", xlab="Кількість кластерів", ylab="Сума квадратів
відстаней всередині кластерів") # побудова графіка
```

Оптимальну кількість кластерів визначають за точкою перегину графіка (рис. 8). У даному випадку малопомітний перегин відповідає трьом кластерам (пригадаємо, що Orange також дає найбільшу оцінку силуета саме для трьох кластерів).

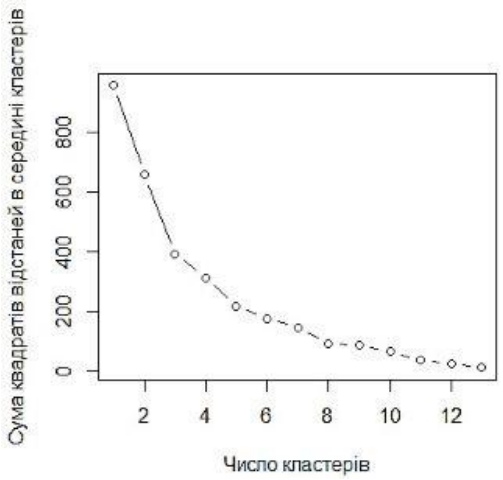

Рис. 8. Графік для визначення оптимальної кількості кластерів

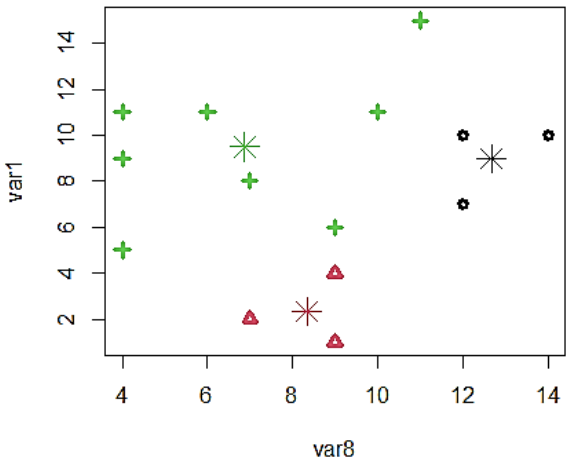

Рис. 9. Діаграма розсіювання

Тепер виконаємо кластерний аналіз методом k-середніх для 3-х кластерів за допомогою функції kmeans:

```
> km <- kmeans(liri, 3)
```

Щоб переконатися у результатах аналізу визначимо середні значення усіх аналізованих параметрів у кожному з кластерів:

```
> aggregate(liri,by=list(km$cluster),FUN=mean)
     Group.1 var1    var2    var3    var4    var5    var6    var7    var8
1    1       9.000   7.333   7.667   11.0    13.000  12.667  12.333  12.667
2    2       2.333   3.667   3.000   3.0     9.333   9.00    8.667   8.333
3    3       9.500   5.875   8.500   8.5     6.875   6.750   9.750   6.875
```

З отриманої таблиці видно різницю між записами в різних кластерах. Тепер зафіксуємо приналежність до кластера для кожного об’єкта, додавши стовпець до фрейма:

```
> liri <- data.frame(liri, km$cluster)
```

Відтак набір даних матиме такий вигляд:

```
> head(liri)
     var1 var2 var3 var4 var5 var6 var7 var8 km.cluster
s_1  2    2    3    3    10   10   8    7    2
s_2  15   9    8    8    8    7    13   11   3
s_3  9    5    9    9    8    3    9    4    3
s_4  5    5    9    9    3    9    8    4    3
s_5  8    6    10   10   7    7    9    7    3
s_6  1    4    2    2    9    7    8    9    2
```

Побудуємо графіки парного співвідношення двох характеристик (наприклад, першої та восьмої, як у випадку з Orange). Приналежність до кластерів на графіку відображена кольором (col), формою (pch) та товщиною контура маркерів (lwd), а центри кластерів буде позначено сніжинками (рис. 9):

```
> op <- par(mfrow = c(1,2))
> plot(liri[c("var8", "var1")], pch=km$cluster, col=km$cluster, lwd=3)
> points(km$centers[,c("var8", "var1")], col=1:3, pch=8, cex=2)
```

За допомогою бібліотеки cluster можна побудувати типовий графік для кластерного аналізу. Заштриховані області різного кольору – це поля параметрів об’єктів, які відносяться до різних кластерів (рис. 10).

```
> library(cluster)
> clusplot(liri, km$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```

Для проведення ієрархічної кластеризації в R можна застосувати, наприклад, функцію hclust(d, method = "complete"),

де:

d – матриця відстаней, отримана за допомогою функції dist() чи іншим способом;

method – метод агломерації, що визначається одним із значень “ward.D”, “ward.D2”, “single”, “complete”, “average”, “mcquitty”, “median” або “centroid”;

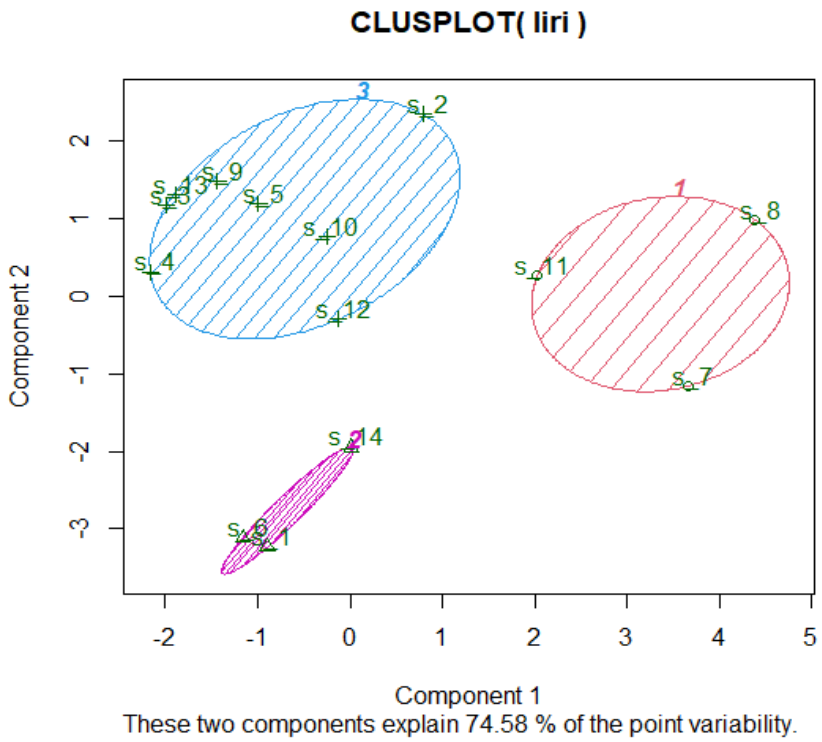

Рис. 10. Діаграма розсіювання із заштрихованими областями

Виконаємо ієрархічний кластерний аналіз для набору даних liri:

```
library(cluster)
d <- dist(scale(liri), method = "euclidean")

par(mfrow = c(4, 1)) #об'єднання чотирьох наступних графіків в одне вікно
par(mar=c(1,1,1,1))  # поля
# Параметр hang=-1 вирівнює мітки
plot(hclust(d, method = "average" ), cex = 0.7, hang = -1)
plot(hclust(d, method = "single" ), cex = 0.7)

res.hc <- hclust(d, method = "complete" )
grp <- cutree(res.hc, k = 4)  #Розрізання дерева на 4 групи
plot(res.hc, cex = 0.7)
rect.hclust(res.hc, k = 4, border = 2:5)

hcd <- as.dendrogram(hclust(d, method = "ward.D2" ))
nodePar <- list(lab.cex = 0.7, pch = c(NA, 19), cex = 0.7, col = "blue")
plot(hcd,  xlab = "Height", nodePar = nodePar, horiz = TRUE, edgePar = list(col
= 2:3, lwd = 2:1))
```

В результаті виконання наведеного скрипта отримаємо такі графіки, як показано на рис. 11 та рис. 12, хоча будувати усі не обов'язково.

Слід зазначити, що бібліотеки R постійно поповнюються, можливості використання розширюються, також існує велика кількість блогів, форумів, присвячених R. Мова непогано документується, і у супровідній документації наведено велику кількість прикладів, наборів даних тощо. Однак, як і у будь-якій мові програмування, в R доводиться багато працювати в консольному режимі, вивчати особливості застосування тих чи інших функцій, підключати додаткові

бібліотеки і т.п., що забирає час, відволікає увагу і зусилля від власне аналізу даних. А от при застосуванні пакетів інтелектуального аналізу даних можна більше уваги приділити форматам представлення даних та інтерпретації отриманих результатів.

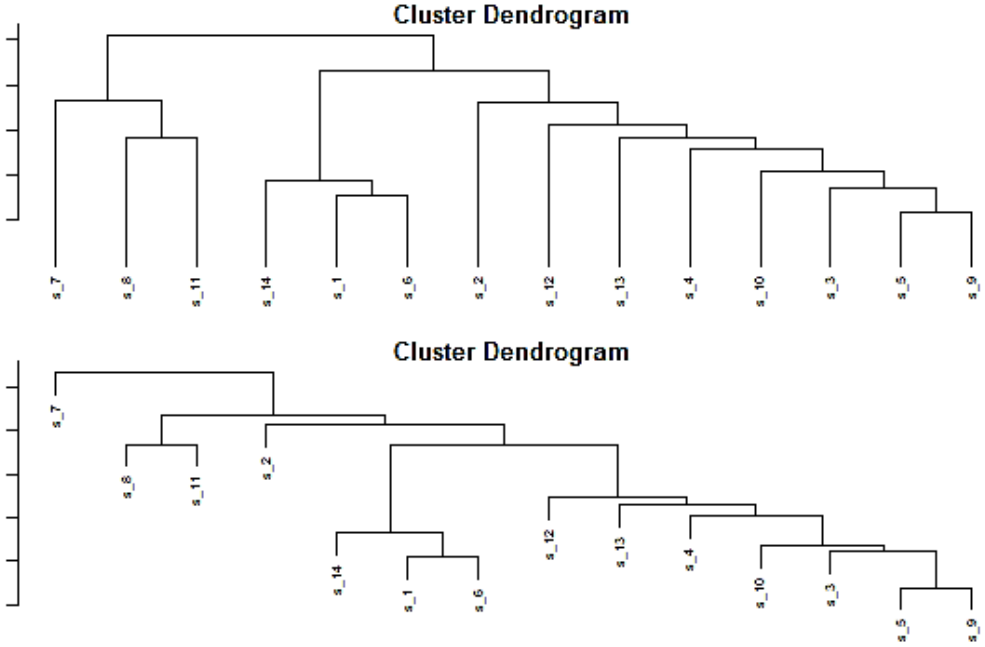

Рис. 11. Дендрограми

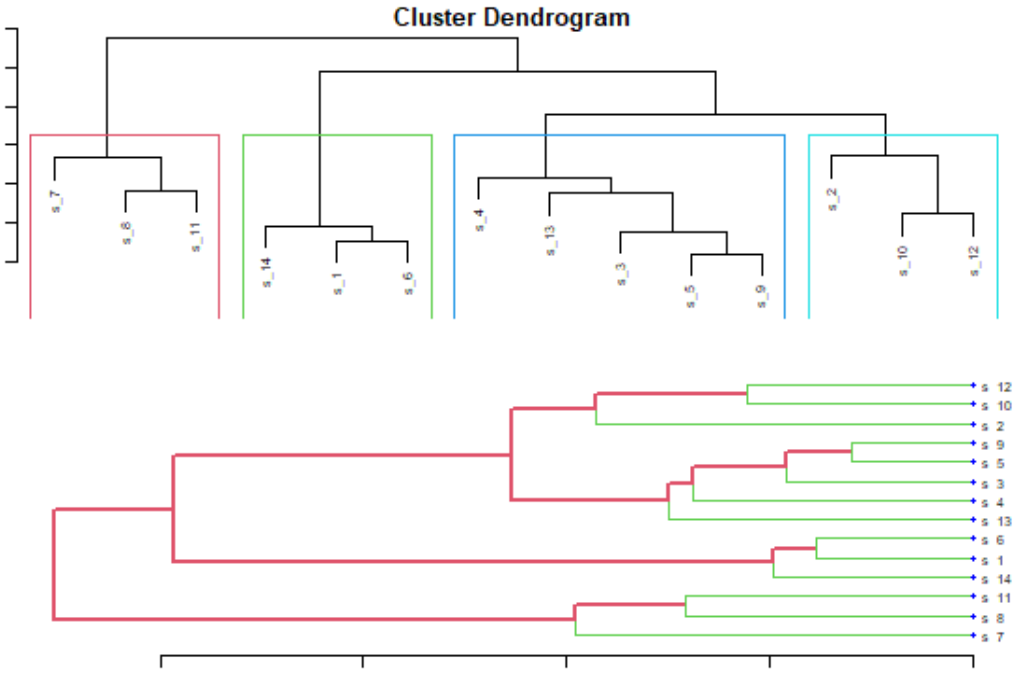

Рис. 12. Дендрограми

Функціонал вільнопоширюваних засобів аналізу даних

Сучасний арсенал методів інтелектуального аналізу даних доволі широкий, що обумовлює різноманітність навчальних програм з даної дисципліни у вищих навчальних закладах. Тим не менше можна виділити інваріантний, спільний для всіх набір методів, куди входять основи

статистичного аналізу (так звана описова статистика), кластерний, факторний, дискримінантний аналіз тощо. Перелік наведених у таблиці засобів включає розглянуті нами (JASP, Orange, R, KNIME, RapidMiner) та такі, що найчастіше згадуються у проаналізованих нами публікаціях, зокрема [2–3], та навчальних програмах.

Як бачимо, табличні процесори, в тому числі MS Excel, на виконання процедур інтелектуального аналізу даних не налаштовані, а от функціонал вільнопоширюваних пакетів забезпечує результати виконання процедур аналізу даних ідентичні до тих, що можна отримати за допомогою таких засобів, як SPSS, RapidMiner тощо. Зазначимо, що асортимент методів інтелектуального аналізу даних наведений перелік не вичерпує.

Результати педагогічного експерименту

За роки викладання курсу інтелектуального аналізу даних у вищому навчальному закладі довелося попрацювати у різних форматах організації навчального процесу та з різним програмним забезпеченням. Це дозволило нам зібрати матеріал для педагогічного експерименту. На підготовчому етапі були розроблені завдання для лабораторних робіт та тестові завдання для перевірки знань.

Таблиця 2

Порівняння аналітичного інструментарію розглянутих засобів

Програмні засоби / Аналітичні процедури	MS Excel	Libre Office Calc	Goggle Sheets	SPSS	RapidMiner	KNIME	Orange	JASP	Python	R
Описова статистика (Descriptives)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
t-тести (T-Tests)	✓	✓	✓	✓	✓	✓		✓	✓	✓
Однофакторний дисперсійний аналіз (ANOVA)	✓	✓	✓	✓	✓	✓		✓	✓	✓
Регресія (Regression)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Частоти (Frequencies)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Розподіли (Distributions)				✓			✓	✓	✓	✓
Факторний аналіз (Factor, PCA – Principal Component Analysis)				✓	PCA	PCA	PCA	✓	✓	✓
Машинне навчання (Machine Learning)				✓	✓		✓	✓	✓	✓
Кластерний аналіз (Clustering)				✓	✓	✓	✓	✓	✓	✓
Дискримінантний аналіз (Discriminant analysis)				✓	✓	✓	✓	✓	✓	✓
Часові ряди (Time Series)				✓	✓	✓	✓	✓	✓	✓
Дерева рішень (Decision trees)				✓	✓	✓	✓	✓	✓	✓
Нейронні мережі (Neural networks)				✓	✓	✓		✓	✓	✓
Пошук асоціативних правил (Association rules – market basket analysis)						✓	✓		✓	✓

Порівняємо навчальні результати студентів двох груп спеціальності 122 Комп'ютерні науки, після вивчення ними теми «Кластерний аналіз». Група КН18б (навчальний рік 2021–2022) виконувала практичні завдання, використовуючи засоби мови R. Група КН19б (навчальний рік 2022–2023) використовувала пакет KNIME. У рамках проведеного пілотного дослідження вдалося забезпечити зменшення впливу сторонніх факторів, оскільки студенти обох груп навчалися за однією методикою (програма, викладач, завдання), виконували один і той самий підсумковий тест тощо.

Після вивчення теоретичних відомостей та виконання лабораторних робіт, які передбачали опанування процедур ієрархічного кластерного аналізу та аналізу k-середніх, студенти відповідали на питання тесту.

Результати розподілилися таким чином (табл. 3, рис. 13):

Таблиця 3

Результати педагогічного експерименту

Навчальний рік	Кількість студентів		
	2021–2022	2022–2023	*Разом*
Група	*кн18б*	*кн19б*	
1–10 балів	4	3	7
11–20 балів	6	4	10
21–30 балів	6	9	15
Разом	16	16	32

Для статистичного порівняння розподілів отриманих балів ми застосували критерій χ^2 (застосування параметричних методів статистичного порівняння до вибірки такого розміру було б некоректним) та отримали: $\chi^2_{\text{емпіричне}}=1{,}143$ ($p=0{,}565$) при $\chi^2_{\text{теоретичне}}=5{,}99$ ($p=0{,}05$), тобто $\chi^2_{\text{емпіричне}} < \chi^2_{\text{теоретичне}}$. Відповідно, гіпотезу H_0 про те, що порівнювані розподіли статистично не відрізняються, відхилити немає підстав.

Рис. 13. Порівняння результатів експериментальних груп

Тим не менше, враховуючи малу чисельність досліджуваної вибірки, робити остаточний висновок про те, що успішність опанування методу аналізу (у даному випадку кластерного аналізу) не залежить від засобу (пакета), який застосовувався у навчальному процесі, передчасно. Бажано провести більш масштабний експеримент.

Висновки. Отже, підсумовуючи усе вищесказане, можна зробити висновок про те, що для вивчення основ інтелектуального аналізу даних у закладі вищої освіти в умовах очного, дистанційного або змішаного навчання можна використовувати безкоштовне програмне забезпечення, якщо його функціонал відповідає цілям навчального процесу.

В рамках нашого дослідження не було виявлено відмінностей у результатах навчання методів інтелектуального аналізу даних (на прикладі кластерного аналізу) при застосуванні як мов програмування (наприклад, R), так і спеціалізованих аналітичних пакетів, хоча для отримання переконливіших результатів експеримент бажано повторити на більш об'ємній вибірці та, можливо, з більшим набором програмних засобів, зокрема поза межами нашого дослідження залишилися такі пакети як Weka, Scikit-learn та інші.

Також у подальших дослідженнях є сенс звернутися до таких актуальних питань як добір та особливості використання вільнопоширюваних засобів аналізу даних для роботи з методами машинного навчання (ML), аналізу текстів (Text Mining) та аналізу великих даних (Big Data Analysis).

Література:

1. Стандарт вищої освіти України першого (бакалаврського) рівня ступеня «бакалавр» за галуззю знань 12 «Інформаційні технології» спеціальністю 122 «Комп'ютерні науки» : МОН України, 2019. URL: https://mon.gov.ua/storage/app/media/vishcha-osvita/zatverdzeni%20standarty/2019/07/12/122-kompyut.nauk.bakalavr-1.pdf.

2. Chakrabarty P., Halder K., Rao P. Tools and Methods of Educational Data Mining: A Review. Easy Chair Preprint №9763, 2023. URL: https://easychair.org/publications/preprint_download/zQDg.

3. Dol S. M., Jawandhiya P. M. A Review of Data Mining in Education Sector. *Journal of Engineering Education Transformations*. 2023. no. 36 (Special Issue 2), pp. 13–22. https://doi.org/10.16920/jeet/2023/v36is2/23003.

4. Shrivastava A., Jain J. K., Chauhan D. "Literature Review on Tools & Applications of Data Mining. *International Journal of Computer Sciences and Engineering*, 2023. vol.11, Issue 4, pp. 46–54. https://doi.org/10.26438/ijcse/v11i4.4654. URL: https://www.ijcseonline.org/pdf_paper_view.php?paper_id=5560&8-IJCSE-09093.pdf.

5. Altalhi A. H., Luna J. M., Vallejo M. A., Ventura S. Evaluation and comparison of open source software suites for data mining and knowledge discovery. *WIREs Data Mining and Knowledge Discovery*, 2017. Vol. 7, Issue 3. https://doi.org/10.1002/widm.1204.

6. Pawar S., Stanam A. Scalable, Reliable and Robust Data Mining Infrastructures/. *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, London, UK, 2020. pp. 123–125. https://doi.org/10.1109/WorldS450073.2020.9210388.

7. Almeida P., Gruenwald L., Bernardino J. Evaluating Open Source Data Mining Tools for Business. *Proceedings of the 5th International* Conference *on Data Management Technologies and Applications – DATA*, 2016. pp. 87–94. http://dx.doi.org/10.5220/0005939900870094.

8. Almeida P., Bernardino J. A Survey on Open Source Data Mining Tools for SMEs. *In: Rocha, Á., Correia, A., Adeli, H., Reis, L., Mendonça Teixeira, M. (eds) New Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing*, Springer, Cham, 2016. vol 444, pp. 253–262. https://doi.org/10.1007/978-3-319-31232-3_24.

9. Özkan S. B., Apaydin S. M. F., Özkan Y., Düzdar I. Comparison of Open Source Data Mining Tools: Naive Bayes Algorithm Example. *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, Istanbul, Turkey, 2019. pp. 1–4. https://doi.org/10.1109/EBBT.2019.8741664.

10. Jovic A., Brkic K., Bogunovic N. An overview of free software tools for general data mining. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* Opatija, Croatia, 2014. pp. 1112–1117. https://doi.org/10.1109/MIPRO.2014.6859735.

11. Al-Odan H. A., Al-Daraiseh A. A. Open Source Data Mining tools. *2015 International Conference on Electrical and Information Technologies (ICEIT)*, Marrakech, Morocco, 2015, pp. 369–374. https://doi.org/10.1109/EITech.2015.7162956.

12. Журан О. А., Донченко К. В. Методи та засоби інтелектуальної обробки інформації. *Інформатика. Культура. Технології: матеріали VIII-ї Міжнародної науково-практичної конференції*, Одеса, Україна, 2021, с. 14–16. URL: http://dspace.op.edu.ua/jspui/bitstream/123456789/12104/1/%D0%86%D0%9A%D0%A2-2021%20%20%D1%81%D0%B1%D0%BE%D1%80%D0%BA%D0%B0%203-14-16.pdf.

13. Лупан І. В. , Авраменко О. В., Акбаш К. С. Комп'ютерні статистичні пакети: навчально-методичний посібник. Кіровоград, Україна: «КОД», 2015. 236 с. URL: https://dspace.cusu.edu.ua/server/api/core/bitstreams/37868982-7a62-4c67-a0c6-acf17c99b48c/content.

14. Лупан І. В. Інтелектуальний аналіз даних Data Mining: навчально-методичний посібник. Кропивницький, Україна: М. А. Піскова, 2022. 112 с. URL: https://dspace.cusu.edu.ua/server/api/core/bitstreams/9df9df5f-ff91-4d35-8497-9a8ac98de872/content.

15. Талах Т., Талах В. Використання функцій Excel в аналітичних дослідженнях та в економічній аналітиці". *Економіка та суспільство*, №50, 2023. http://doi.org/10.32782/2524-0072/2023-50-58.

16. Data analytics and AI platform | Altair RapidMiner. URL: http://altair.com/altair-rapidminer.

17. KNIME Analytics Platform. URL: https://www.knime.com/knime-analytics-platform.

18. Orange Data Mining. URL: http://orangedatamining.com.

19. JASP. A fresh way to do statistics. URL: http://jasp-stats.org.

References:

1. Ministry of Education and Science of Ukraine (2019). Standart vyshchoi osvity Ukrainy pershoho (bakalavrskoho) rivnia stupenia "bakalavr" za haluzziu znan 12 "Informatsiini tekhnolohii" spetsialnistiu 122 "Kompiuterni nauky" [The standard of higher education of Ukraine of the first (bachelor) level of the

"bachelor" degree in the field of knowledge 12 "Information technologies" specialty 122 "Computer science"]. Kyiv: Ministry of Education and Science of Ukraine. Retrieved from: https://mon.gov.ua/storage/app/media/vishcha-osvita/zatverdzeni%20standarty/2019/07/12/122-kompyut.nauk.bakalavr-1.pdf [in Ukrainian].

2. Chakrabarty, P., Halder, K., & Rao, P. (2023). Tools and Methods of Educational Data Mining: A Review. *Easy Chair Preprint* №9763. Retrieved from: https://easychair.org/publications/preprint_download/zQDg [in English].

3. Dol, S. M., & Jawandhiya, P. M. (2023). A review of data mining in education sector. *Journal of Engineering Education/Journal of Engineering Education Transformations/Journal of Engineering Education Transformation*, *36*(S2), 13–22. https://doi.org/10.16920/jeet/2023/v36is2/23003 [in English].

4. Shrivastava, A., Jain, J. K., & Chauhan, D. (2023). Literature Review on Tools & Applications of Data Mining. *International Journal of Computer Sciences and Engineering*, 11(4), 46–54. Retrieved from: https://www.ijcseonline.org/pdf_paper_view.php?paper_id=5560&8-IJCSE-09093.pdf [in English].

5. Altalhi, A. H., Luna, J. M., Vallejo, M. A., & Ventura, S. (2017b). Evaluation and comparison of open source software suites for data mining and knowledge discovery. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, *7*(3). https://doi.org/10.1002/widm.1204 [in English].

6. Pawar, S., & Stanam, A. (2020). Scalable, reliable and robust data mining infrastructures. *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 123–125. https://doi.org/10.1109/worlds450073.2020.9210388 in English].

7. Almeida, P., Gruenwald, L., & Bernardino J. (2016) Evaluating Open Source Data Mining Tools for Business. *In Proceedings of the 5th International* Conference *on Data Management Technologies and Applications – DATA*, 87–94. http://dx.doi.org/10.5220/0005939900870094 [in English].

8. Almeida, P., & Bernardino, J. (2016). A survey on open source data mining tools for SMEs. *In Advances in intelligent systems and computing*, 253–262. https://doi.org/10.1007/978-3-319-31232-3_24 [in English].

9. Özkan, S. B., Apaydin, S. M. F., Özkan, Y., & Düzdar. (2019). Comparison of Open Source Data Mining Tools: Naive Bayes Algorithm Example. *In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, Istanbul, Turkey, 1–4. https://doi.org/10.1109/EBBT.2019.8741664 [in English].

10. Jovic, A., Brkic, K., & Bogunovic, N. (2014). An overview of free software tools for general data mining. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1112–1117. https://doi.org/10.1109/MIPRO.2014.6859735 [in English].

11. Al-Odan, H. A., & Al-Daraiseh, A. A. (2015). Open Source Data Mining tools. *2015 International Conference on Electrical and Information Technologies (ICEIT)*, P. 369–374. https://doi.org/10.1109/eitech.2015.7162956 [in English].

12. Zhuran, O. A., & Donchenko, K. V. (2021). Metody ta zasoby intelektualnoi obrobky informacii [Methods and means of intellectual processing of information]. *International Scientific and Practical Conf. "Informatics. Culture. Technology"*, 14–16. Retrieved from: http://dspace.op.edu.ua/jspui/bitstream/123456789/12104/1/%D0%86%D0%9A%D0%A2-2021%20%20%D1%81%D0%B1%D0%BE%D1%80%D0%BA%D0%B0%203-14-16.pdf [in Ukrainian].

13. Lupan, I. V., Avramenko, O. V., & Akbash, K. S. (2015). Computerni statystychni pakety: navchalno-metodychnyi posibnyk [Computer Statistical Packages: Tutorial]. KOD, 236 p. Retrieved from: https://dspace.cusu.edu.ua/server/api/core/bitstreams/37868982-7a62-4c67-a0c6-acf17c99b48c/content [in Ukrainian].

14. Lupan, I. V. (2022). Intelectualnyi analis danyh: navchalno-metodychnyi posibnyk [Data Mining: Tutorial]. M. A. Piskova, 112 p. Retrieved from: https://dspace.cusu.edu.ua/server/api/core/bitstreams/9df9df5f-ff91-4d35-8497-9a8ac98de872/content [in Ukrainian].

15. Talakh, T., & Talakh, V. (2023). Vykorystannia funkcii Excel v analitychnykh doslidzenniakh ta v ekonomichnii analityci [Using Excel Functions in Analytical Research and Economic Analytics]. *Ekonomika ta suspilstvo*, 50. http://doi.org/10.32782/2524-0072/2023-50-58 [in Ukrainian].

16. Data analytics and AI platform | Altair RapidMiner. Retrieved from: http://altair.com/altair-rapidminer [in English].

17. KNIME Analytics Platform. Retrieved from: https://www.knime.com/knime-analytics-platform [in English].

18. Orange Data Mining. Retrieved from: http://orangedatamining.com [in English].

19. JASP. A fresh way to do statistics. Retrieved from: http://jasp-stats.org [in English].