Методичні вказівки

до виконання контрольних робіт

з дисципліни

«Інтелектуальний аналіз даних»

для студентів заочної форми навчання

спеціальності 122 – «Комп'ютерні науки»

Міністерство освіти і науки України
Вінницький національний технічний університет

Методичні вказівки

до виконання контрольних робіт
з дисципліни

«Інтелектуальний аналіз даних»

для студентів заочної форми навчання
спеціальності 122 – «Комп'ютерні науки»

Вінниця
ВНТУ
2021

Рекомендовано до друку Методичною радою Вінницького національного технічного університету Міністерства освіти і науки України (протокол № 4 від 23.12.2020 р.)

Рецензенти:

Т. Б. Мартинюк, докторка технічних наук, професорка

В. П. Майданюк, кандидат технічних наук, доцент

Методичні вказівки до виконання контрольних робіт з дисципліни «Інтелектуальний аналіз даних» для студентів заочної форми навчання спеціальності 122 – «Комп'ютерні науки» / Уклад. В. І. Месюра, Я. В. Іванчук, О. К. Колесницький. – Вінниця : ВНТУ, 2021. – 42 с.

У методичних вказівках наведено основні теоретичні дані, приклади розв'язання типових задач, вимоги до структури та захисту контрольної роботи, перелік питань з диференційованого заліку з дисципліни «Інтелектуальний аналіз даних» та рекомендовані літературні джерела. Методичні вказівки розроблено відповідно до плану кафедри та робочої програми дисципліни «Інтелектуальний аналіз даних».

ЗМІСТ

ВСТУП

У найзагальнішому сенсі аналіз даних є процесом дослідження, фільтрації, перетворення й моделювання даних з метою отримання корисної інформації.

До початку 90-х років минулого століття роль основного інструменту аналізу даних виконувала математична (прикладна) статистика, що далеко не завжди забезпечувало його високу ефективність. Головною причиною цього була концепція усереднення за вибіркою, яка не рідко вела до операцій над фіктивними величинами (типу середньої температури пацієнтів у лікарні, середньої висоти будинку на вулиці, що складається з палаців і халуп і т. п.). Методи математичної статистики виявилися корисними, головним чином, для перевірки заздалегідь сформульованих гіпотез (verification-driven data mining) і для «грубого» розвідувального аналізу, що становить основу оперативної аналітичної обробки даних (online analytical processing, OLAP).

Технологічний прорив в галузі записування та зберігання даних наприкінці XX століття обрушив на людство колосальні потоки «сирих» даних (інформаційної руди) в найрізноманітніших галузях. Діяльність будь-якого підприємства або організації (комерційної, виробничої, медичної, наукової і т. д.) стала супроводжуватися реєстрацією та записуванням усіх подробиць її діяльності. Накопичена людством інформація перетворила світ на цифровий простір, обсяг якого у 2020 р. досягнув 46 зеттабайтів. Щосекунди на одну людини планети створюється 1,7 мегабайтів нової інформації.

Традиційна математична статистика відверто спасувала перед обличчям проблем, що виникли. Виявлення закономірностей в гігантських обсягах «сирої» інформації, що несе масу неточних, неповних (з пропусками), суперечливих, різнорідних, непрямих даних, потребувала величезних інтелектуальних зусиль.

Сучасна концепція аналізу даних отримала назву Data Mining, що перекладається як «видобуток» або «розкопка даних». Нерідко поруч з Data Mining зустрічаються означення «виявлення знань в базах даних» (knowledge discovery in databases) і «інтелектуальний аналіз даних», які можна вважати синонімами Data Mining, хоча кожне з них концентрує увагу на певних аспектах загального процесу інтелектуального аналізу даних.

Data Mining є процесом виявлення в «сирих» даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності [1]. В основі Data Mining лежить математичний апарат, який виник і розвивається на базі досягнень прикладної статистики, розпізнавання образів, машинного навчання, методів штучного інтелекту, теорії баз даних і т. ін. (рис. 1). Вибір методу часто залежить від типу наявних даних і від того, яку інформацію намагаються отримати.

За основу Data Mining покладено концепцію шаблонів (шаблонів), що відображають фрагменти багатоаспектних взаємовідношень у даних.

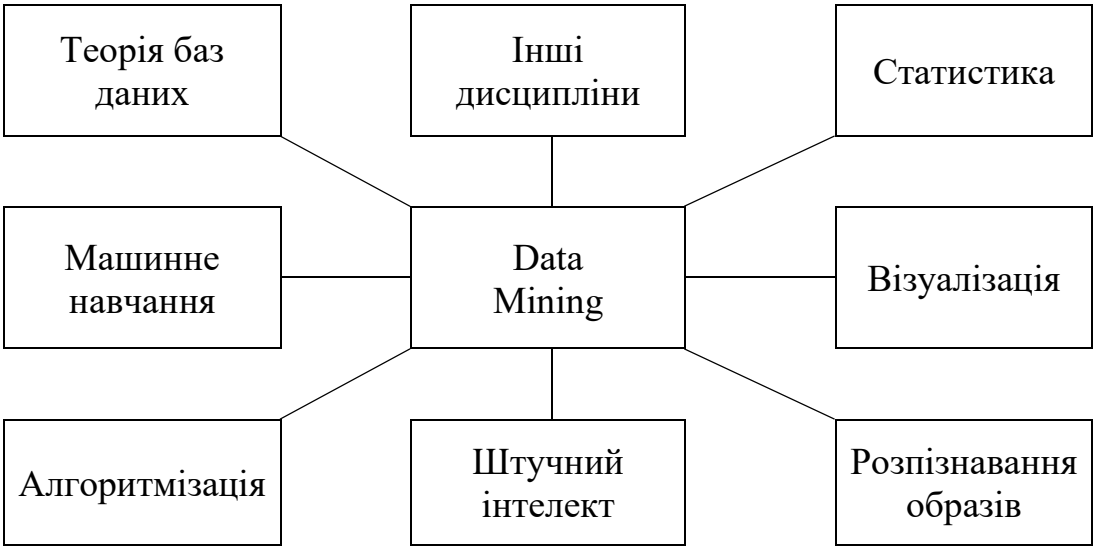

Рисунок 1 – Основні галузі знань, на яких базується Data Mining

Ці шаблони подають закономірності, властиві підвибіркам даних, які можуть бути компактно вираженими в зрозумілій людині формі. Пошук шаблонів здійснюється методами, що не обмежені рамками попередніх припущень про структуру вибірки та вид розподілу значень аналізованих показників.

Приклади завдань такого пошуку шаблонів в Data Mining наведено в таблиці 1.

Таблиця 1 – Приклади формулювань задач при OLAP і Data Mining

OLAP	ІАД
Якими є середні показники травматизму для тих, хто палить і не палить?	Які фактори найчастіше призводять до травматизму?
Якими є середні суми телефонних рахунків наявних клієнтів порівняно з рахунками тих клієнтів, які відмовилися від послуг телефонної компанії?	Які характеристики відрізняють клієнтів, які, імовірно, збираються відмовитись від послуг телефонної компанії?
Якою є середня величина щоденних покупок по вкраденій та не вкраденій кредитній картці?	Які схеми покупок є характерними для випадків шахрайства з кредитними картками?

Різниця між аналізом даних і інтелектуальним аналізом даних полягає в тому, що аналіз даних використовується для перевірки моделей та гіпотез про набір даних, наприклад, для аналізу ефективності маркетингової кампанії, незалежно від кількості даних. На відміну від цього, інтелектуальний аналіз даних використовує машинне навчання і статистичні моделі для виявлення закономірностей, прихованих у великому обсязі даних.

Найбільш поширена методологія інтелектуального аналізу даних CRISP-DM (CRoss Industry Standard Process for Data Mining) складається з таких етапів:

1. Розуміння потреб користувача: визначення цілей і вимог з боку користувача для постановки задачі інтелектуального аналізу даних і формування плану досягнення цілей.

2. Розуміння даних: виявити проблеми щодо якості даних, зрозуміти, що за дані наявні, спробувати відшукати цікаві набори даних або сформувати гіпотези про наявність прихованих закономірностей в даних.

3. Підготовка даних: формування з різнорідних і різноформатних «сирих» даних наборів даних, придатних для моделювання.

4. Моделювання: побудова різноманітних моделей та налаштування їхніх параметрів на оптимальні значення.

5. Оцінення: отримання кількісних оцінок якості побудованої моделі для впевненості у досягненні всіх поставлених бізнес-цілей.

6. Розгортання: складання фінального звіту або автоматизація процесу аналізу даних для вирішення бізнес-завдань, для пояснення клієнту того, що саме йому потрібно зробити для того, щоб ефективно використовувати отримані моделі.

Фактична завданням інтелектуального аналізу даних є напів-автоматичний або автоматичний аналіз великих обсягів даних для вилучення раніше невідомих цікавих шаблонів, таких як групи записів даних (кластерний аналіз), незвичайних записів (виявлення аномалій) і залежностей (інтелектуальний аналіз асоціативних правил, послідовний аналіз шаблонів). Ні збір даних, ні підготовка даних, ні інтерпретація результатів і подання звітів не є частиною етапу інтелектуального аналізу даних, але відносяться до загального процесу KDD як додаткові етапи [3].

Інтелектуальний аналіз даних охоплює шість загальних класів задач:

1. Виявлення аномалій (викидів/змін/відхилень) даних, які можуть бути цікаві або можуть бути помилками даних, що потребують подальшого вивчення.

2. Вивчення правил асоціації – пошук взаємозв'язків між змінними. Наприклад, на основі зібраних даних про купівельні звички клієнтів супермаркет може побудувати асоціативні правила одночасної купівлі певних продуктів і використати цю інформацію в маркетингових цілях.

3. Кластеризація – виявлення так чи інакше «схожих» груп і структур в даних.

4. Класифікація – узагальнення відомої структури для застосування до нових даних. Наприклад, повідомлення електронної пошти спробувати класифікувати як «законне» або як «спам».

5. Регресія – пошук функції оцінювання відношень між даними або наборами даних, яка моделює дані з найменшою помилкою.

6. Узагальнення – забезпечення більш компактного подання набору даних, включно з візуалізацією і створенням звітів.

1 ЗМІСТ ДИСЦИПЛІНИ

Змістовний модуль 1. Основи інтелектуального аналізу даних

Тема 1. Основи інтелектуального аналізу даних

Методи первісної обробки даних. Методи дослідження структури даних: візуалізація та автоматичне групування даних.

Тема 2. Методи використання навчальної інформації

Кореляційний і регресійний аналіз даних. Множинний регресійний аналіз. Лінійна множинна регресійна модель. Перевірка адекватності моделі. Нелінійне оцінювання параметрів.

Змістовний модуль 2. Методи багатовимірного розвідувального аналізу

Тема 3. Методи кластерного аналізу

Кластерний аналіз. Ієрархічна та секційна кластеризація. Методи кластеризації: процедура Мак-Кіна, метод k-методів, сітчасті методи. Растрова кластеризація об'єктів. Лінійний дискримінантний аналіз. Побудова канонічних та класифікаційних функцій.

Тема 4. Методи класифікації та прогнозування

Дерева рішень. Методи опорних векторів, «найближчого сусіда», Байєса. Аналіз багатовимірних угруповань.

Тема 5. Статистична обробка даних

Статистична обробка часових рядів і прогнозування. Класифікація об'єктів у випадку невідомих розподілень даних. Методи оцінювання помилок класифікації.

Змістовий модуль 3. Методи пошуку шаблонів даних

Тема 6. Методи пошуку асоціативних правил

Асоціаційні правила. Послідовне відображення шаблонів даних. Метод Apriori, побудова FP-дерев пошуку шаблонів даних.

Тема 7. Технологія аналітичної обробки даних в реальному часі

Min-max асоціації у базах даних. Побудова hash-дерев. Розробка OLAP-кубів під час аналізу багатовимірних даних у великих БД. Способи та методи візуального відображення даних.

Змістовний модуль 4. OLAP і Data Mining

Тема 8. Методи і стадії Data Mining

Методи, стадії, задачі Data Mining. Упровадження Data Mining, OLAP і сховищ даних у СППР.

Тема 9. Процес Data Mining

Процес Data Mining. Стандарти Data Mining. Інструменти Data Mining.

2 ЗАВДАННЯ ДО КОНТРОЛЬНОЇ РОБОТИ

Загальне завдання до контрольної роботи складається з двох теоретичних та чотирьох практичних завдань.

Теоретичне питання вибирається з переліку питань для заліку з номером

$$T = x + p,$$

де x – номер студента в групі,

p – число, яке окремо задається викладачем для кожної групи.

Практичні завдання переважно мають формулюватись відповідно до теми бакалаврської дипломної роботи і відноситись до розв’язання таких класів задач:

- лінійна регресія;
- наївний байєсовський класифікатор;
- дерева рішень;
- карти Кохонена;
- кластеризація;
- асоціативні правила;
- генетичні алгоритми.

Типові приклади розв’язання практичних завдань наведено у розділі 3 цих методичних вказівок.

3 ПРИКЛАДИ РОЗВ'ЯЗАННЯ ТИПОВИХ ПРАКТИЧНИХ ЗАВДАНЬ

3.1 Лінійна регресія

Модель регресійного аналізу використовується для прогнозування значення однієї залежної змінної, виходячи з відомих значень незалежних змінних.

Завдання. Надано статистичні дані про одноденний середньодушовий прожитковий мінімум працездатної особи та її середньоденну зарплату (табл. 2) [2].

Таблиця 2 – Статистичні дані про одноденний середньодушовий прожитковий мінімум працездатної особи та її середньоденну зарплату

ID регіону	Середньодушовий прожитковий мінімум працездатної особи за день	Середньоденна зарплата
1	75	133
2	78	125
3	81	129
4	93	153
5	86	140
6	77	135
7	85	135
8	77	132
9	89	161
10	95	159
11	72	120
12	115	160

1. Побудувати лінійне рівняння парної регресії y від x.

2. Розрахувати лінійний коефіцієнт парної кореляції і середню помилку апроксимації.

3. Оцінити статистичну значущість параметрів регресії і кореляції за допомогою критеріїв Фішера і Стьюдента.

4. Виконати прогноз заробітної плати при прогнозному значенні середньодушового прожиткового мінімуму, що становить 107% від середнього рівня.

5. Оцінити точність прогнозу, розрахувавши помилку прогнозу і його довірчий інтервал.

6. На одному графіку побудувати вихідні дані і теоретичну пряму.

Розв'язання. Обчислимо характеристики випадкових величин X і Y (вибіркове середнє і вибіркове середнє квадратичне відхилення).

1. Для розрахунку параметрів рівняння лінійної регресії побудуємо розрахункову таблицю (табл. 3).

Таблиця 3 – Розрахункові дані для визначення параметрів рівняння лінійної регресії

№	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$\frac{\lvert(y_i - \hat{y}_i\rvert}{y_i} \times 100$
1	75	133	9975	5625	17689	129.808	3.192	10.189	2.400
2	78	125	9750	6084	15625	132.844	-7.844	61.528	6.275
3	81	129	10449	6561	16641	135.88	-6.88	47.334	5.333
4	93	153	14229	8649	23409	148.024	4.976	24.761	3.252
5	86	140	12040	7396	19600	140.94	-0.94	0.884	0.671
6	77	135	10395	5929	18225	131.832	3.168	10.036	2.347
7	85	135	11475	7225	18225	139.928	-4.928	24.285	3.650
8	77	132	10164	5929	17424	131.832	0.168	0.028	0.127
9	89	161	14329	7921	25921	143.976	17.024	289.817	10.574
10	95	159	15105	9025	25281	150.048	8.952	80.138	5.630
11	72	120	8640	5184	14400	126.772	-6.772	45.860	5.643
12	115	160	18400	13225	25600	170.288	-10.288	105.843	6.430
Усього:	1023	1682	144951	88753	238040			700.703	52.334
Середнє значення	85.250	140.167	12079.250	7396.083	19836.667				4.361
σ	11.337	13.783							
σ^2	128.521	189.972							

$a_1 = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{12079{,}5 - 140{,}167 \cdot 85{,}25}{7396{,}083 - 85{,}25^2} = 1{,}012$..

$a_0 = \bar{y} - a_1 \bar{x} = 140{,}167 \cdot 85{,}25 = 53{,}894$.

Отже, рівняння лінійної регресії набуде вигляду:

$$y = 53{,}894 + 1{,}012x.$$

Це означає, що за зростання середньодушового прожиткового мінімуму на 1 грн середньоденна зарплата зростатиме в середньому на 1,012 грн.

2. З використанням коефіцієнта кореляції оцінимо щільність лінійного зв'язку

$$r_{xy} = a_1 \frac{\sigma_x}{\sigma_y} = 1{,}012 \cdot \frac{11{,}337}{13{,}783} = 0{,}832.$$

Коефіцієнт детермінації:

$$r_{xy}^2 = 0{,}832^2 = 0{,}692.$$

Отже, 69,2% варіації зарплати (y) обумовлюється варіацією фактора x – середньодушового прожиткового мінімуму.

Шляхом обчислення середньої помилки апроксимації, оцінимо якість побудованої моделі:

$$\bar{A}=\frac{1}{n}=\sum\frac{|y_i-\hat{y}_i|}{|y_i}\cdot 100=\frac{52{,}334}{12}=4{,}361\%.$$

Якість побудованої моделі оцінюються як гарна, оскільки середня помилка апроксимації не перевищує 8–10%.

3. Розрахуємо F-критерій:

$$F_{\text{факт}}=\frac{r_{xy}^2}{1-r_{xy}^2}(n\text{-}2)=\frac{0{,}692}{1-0{,}692}=22{,}5.$$

За таблицею F-розподілу Фішера-Снедекора, при рівні значущості $\alpha = 0{,}05$ і кількості ступенів свободи k_1=1 і k_2=12-2=10, критичне значення становитиме:

$$F_{\text{таб}}=4{,}96;\quad F_{\text{факт}}>F_{\text{таб}}.$$

H_0 – гіпотеза про статистичну незначимість рівняння регресії відхиляється.

Оцінення статистичної значущості параметрів регресії проведемо за допомогою t-статистики Стьюдента і шляхом довірчого інтервалу кожного з показників.

Висуваємо гіпотезу H_0 про статистичну незначимість відмінності показників від нуля:

$$a=b=r_{xy}=0.$$

$t_{\text{табл}}$ для кількості ступенів свободи $df = n - 2 = 12 - 2 = 10$ і $a = 0{,}05$ становитиме 2,23.

Визначимо випадкові помилки m_a, m_b, $m_{r_{xy}}$:

$$S_{\text{ост}}=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{700{,}703}{12-2}}=8{,}371;$$

$$m_{a_0}=\frac{S_{\text{ост}}}{n}\cdot\sqrt{\frac{\overline{x^2}}{\overline{x^2}-\bar{x}^2}}=8{,}371\cdot\frac{\sqrt{7396{,}083}}{12\cdot 11{,}377}=5{,}292;$$

$$m_{a_1} = \frac{S_{\text{ост}}}{\sqrt{n(\overline{x^2} - \bar{x}^2)}} = \frac{8{,}371}{11{,}377 \cdot \sqrt{12}} = 0{,}213;$$

$$m_{r_{xy}} = \cdot \sqrt{\frac{1 - r_{xy}^2}{n-2}} = \sqrt{\frac{1-0{,}692}{12-2}} = 0\text{б}176.$$

Тоді:

$$t_{a_0} = \frac{a_0}{m_a} = \frac{53{,}894}{5{,}292} = 10{,}184;$$

$$|t_{a_0}| = 10{,}184 > t_{\text{табл}} = 2{,}23.$$

Фактичне значення є вищим за табличне значення *t*-статистики. Нульова гіпотеза відхиляється, оскільки a_0 не випадково відрізняється від нуля, а статистично значимо.

$$t_{a_1} = \frac{a_1}{m_b} = \frac{1{,}012}{0{,}213} = 4{,}751;$$

$$|t_{a_1}| = 4{,}751 > t_{\text{табл}} = 2{,}23.$$

Фактичне значення є вищим за табличне значення *t*-статистики. Нульова гіпотеза відхиляється, оскільки a_1 не випадково відрізняється від нуля, а статистично значимо.

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0{,}832}{0{,}176} = 4{,}727;$$

$$|t_{r_{xy}}| = 4{,}727 > t_{\text{табл}} = 2{,}23.$$

Фактичне значення є вищим за табличне значення *t*-статистики. Нульова гіпотеза відхиляється, оскільки $|t_{r_{xy}}$ не випадково відрізняється від нуля, а статистично значимо.

Розрахуємо довірчі інтервали для параметрів регресії a_0 і a_1. Для цього визначимо граничну помилку для кожного показника:

$$\Delta_{a_0} = t_{\text{табл}} \cdot m_{a_0} = 2{,}23 \cdot 5{,}392 = 12{,}024;$$

$$\Delta_{a_1} = t_{\text{табл}} \cdot m_{a_1} = 2{,}23 \cdot 0{,}213 = 0{,}475.$$

Довірчі інтервали:

$$\gamma_{a_0} = a_0 \pm \Delta_{a_0} = 53{,}894 \pm 12{,}024 \text{ або } 41{,}87 < a_0 < 65{,}918;$$
$$\gamma_{a_1} = a_1 \pm \Delta_{a_1} = 1{,}012 \pm 0{,}475 \text{ або } 0{,}537 < a_1 < 1{,}487.$$

4. Отримані оцінки рівняння регресії дозволяють використовувати його для прогнозу. Якщо прогнозне значення прожиткового мінімуму становитиме:

$$x_p = \bar{x} \cdot 1{,}07 = 85{,}25 \cdot 1{,}07 = 91{,}218 \text{ грн},$$

то прогнозне значення середньоденної зарплати буде:

$$y_p = 53{,}894 + 1{,}012 \cdot 91{,}218 = 146{,}207 \text{ грн}.$$

5. Помилка прогнозу становитиме:

$$m_p = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n\sigma_x^2}} = 9{,}521 \cdot \sqrt{1 + \frac{1}{12} + \frac{(91{,}218 - 85{,}25)^2}{12 \cdot 128{,}521}} = 10{,}015 \text{ грн}.$$

Гранична помилка прогнозу, яка не буде перевищена у 95% випадків, становитиме:

$$\Delta = t_{\text{набл}} \cdot m_p = 2{,}23 \cdot 10{,}015 = 22{,}334 \text{ грн}.$$

Довірчий інтервал прогнозу:

$$(146{,}207 - 22{,}334;\ 146{,}207 + 22{,}334) = (123{,}873;\ 168{,}541) \text{ грн}.$$

6. Побудуємо вихідні дані і теоретичну пряму (рис. 2).

3.2 Класифікація

Класифікація визначає належність конкретного об’єкта певній групі об’єктів залежно від значень його параметрів.

3.2.1 Наївний Байєсовський класифікатор

Завдання. Нехай відомо, що 0,8% людей генетично схильні до захворювання на рак. Існуючі тести не є ідеальними і повертають правильний позитивний результат в 98 % випадків, якщо схильність наявна, і правильний негативний результат в 97 % випадків, коли схильності немає:

– розрахуйте апостеріорну ймовірність того, що при отриманні позитивного результату тесту пацієнт дійсно схильний до захворювання;

– знаючи про неідеальність тесту, пацієнт проходить другий тест, який вважається незалежним від першого. Розрахуйте апостеріорні ймовірності для наявності і відсутності схильності до захворювання особи, у разі, якщо другий тест дав позитивний результат.

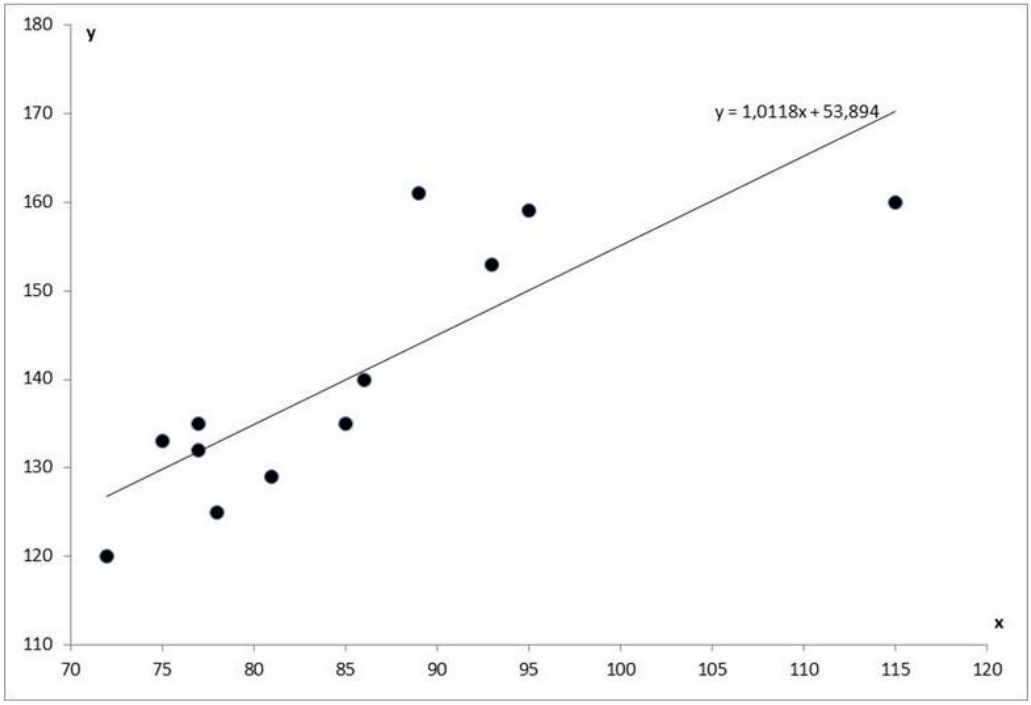

Рисунок 2 – Вихідні дані та теоретична пряма лінійної регресії

Розв'язання. Для розв'язання задачі скористаємось правилом:
ЯКЩО гіпотеза H = істина, ТО свідоцтво E спостерігається з імовірністю P.

Визначимо імовірність захворювання на рак H особи, яка отримала позитивний результат тесту E, що має імовірність P:

$$p(H/E) = \frac{p(E/H)\times p(H)}{p(E/H)\times p(H)+\mathrm{p}(E/\neg H)\times p(\neg H)}.$$

Маємо:
– апріорна імовірність захворювання на рак $p(H)$=0,8;
– апріорна імовірність відсутності захворювання:

$$p(\neg H)= 1 - H = 1 - 0{,}008 = 0{,}992;$$

– апріорна імовірність ПОЗитивного результату тесту за наявності раку:

$$\mathrm{P(ПОЗ|рак)} = 0{,}98;$$

– апріорна імовірність НЕГативного результату тесту за наявності раку:

$$p(\mathrm{НЕГ|\ рак}) = 1 - 0{,}98 = 0{,}02;$$

– апріорна імовірність правильного (НЕГативного) результату тесту за відсутності раку:

$$p(\text{НЕГ}| \neg \text{рак}) = 0{,}97;$$

– апріорна імовірність ПОЗитивного результату тесту при відсутності раку:

$$p(\text{ПОЗ}| \neg \text{рак}) = 1 - 0{,}97 = 0{,}03.$$

Відбулась подія: ПОЗитивний результат тесту
Обчислюємо апостеріорну імовірність:

$$p(\text{рак}| \text{ПОЗ}) = p(\text{ПОЗ}|\text{рак}) \times p(\text{рак}) = 0{,}98 \times 0{,}008 = 0,\ 0078;$$
$$p(\neg \text{рак}| \text{ПОЗ}) = p\ (\text{ПОЗ}| \neg \text{рак}) \times p\ (\neg \text{рак}) = 0{,}03 \times 0{,}992 = 0{,}0298\text{l}.$$

Апостеріорна імовірність захворювання на рак особи, яка отримала позитивний тест:

$$p\ (\text{рак}| \text{ПОЗ}) = 0{,}0078/(0{,}0078 + 0{,}0298) = 0{,}21.$$

Апостеріорна імовірність того, що особа, яка отримала позитивний результат тестування не хворіє на рак:

$$p(\neg \text{рак}| \text{ПОЗ}) = 1 - p\ (\text{рак}| \text{ПОЗ}) = 1 - 0{,}21 = 0{,}79.$$

Визначимо імовірність захворювання на рак H особи, яка повторно отримала позитивний результат тесту E, що має імовірність P:

$$p(H/E_1, E_2) = \frac{p(E_1/H) \times p(E_2/H) \times p(H)}{p(E_1/H) \times p(E_2/H) \times p(H) + \text{p}(E_1/\neg H) \times p(E_2/\neg H) \times p(\neg H))};$$

$$\text{Р(рак}| \text{ПОЗ, ПОЗ)} = \text{Р(ПОЗ}|\text{рак)} \times \text{Р(ПОЗ}|\text{рак)} \times \text{Р(рак)} =$$
$$= 0{,}98 \times 0{,}98 \times 0{,}008 = 0,\ 00768;$$

$$\text{Р}(\neg \text{рак}| \text{ПОЗ, ПОЗ}) = \text{Р(ПОЗ}| \neg \text{рак)} \times \text{Р(ПОЗ}| \neg \text{рак)} \times \text{Р}(\neg \text{рак}) =$$
$$= 0{,}03 \times 0{,}03 \times 0{,}992 = 0{,}0008928.$$

Апостеріорна імовірність захворювання на рак особи, яка отримала два позитивних результати тестування становитиме:

$$\text{Р(рак}| \text{ПОЗ, ПОЗ)} = 0{,}00768/(0{,}00768 + 0{,}0008928) =$$
$$= 0{,}00768/0{,}00857 = 0{,}896.$$

Апостеріорна імовірність того, що особа, яка отримала два позитивних результати тестування, не хворіє на рак:

$$\text{Р}(\neg \text{рак}| \text{ПОЗ,ПОЗ}) = 1 - \text{Р(рак}| \text{ПОЗ,ПОЗ}) = 1 - 0{,}896 = 0{,}104.$$

3.2.2 Класифікатор на основі дерева рішень

Завдання. Для проведення чергової маркетингової компанії відділ маркетингу проводить дослідження із визначення, які групи покупців найбільш схильні до придбання комп'ютера. Наявна інформація (табл. 4) щодо:

- віку (≤ 30 / $31-40$ / >40);
- доходу (низький / середній / високий);
- кредитної історії (гарна / відмінна);
- статусу (студент / не студент);
- історії купівлі комп'ютера за результатами минулих маркетингових компаній.

Таблиця 4 – Начальна вибірка для побудови дерева рішень щодо схильності клієнта до купівлі комп'ютера

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Вік	<=30	<=30	31-40	>40	>40	>40	31-40	<=30	<=30	>40	<=30	31-40	31-40	>40
Дохід	висок	висок	висок	серед	низьк	низьк	низьк	серед	низьк	серед	серед	серед	висок	серед
Студент	ні	ні	ні	ні	так	так	так	ні	так	так	так	ні	так	ні
Кредит	гарна	відм	гарна	гарна	гарна	відм	відм	гарна	гарна	гарна	відм	відм	гарна	відм
Придбання	ні	ні	так	так	так	ні	так	ні	так	так	так	так	так	ні

Ціль дослідження – за наявними характеристиками клієнта за допомогою алгоритму ID3 побудувати дерево рішень, що дозволить виявити ступінь зацікавленості клієнта у придбанні комп'ютера.

Розв'язання. При побудові дерева рішень для вибору атрибуту для розбиття скористуємося критерієм Gain приросту ентропії.

1. Розрахуємо ентропію кореневої вершини дерева рішень:

$$I(S_{ТАК}, S_{НІ}) = I(9,5) = -9/14 \log_2(9/14) - 5/14 \log_2(5/14) = 0.94.$$

2. Розрахуємо приріст ентропії, який забезпечить розбиття дерева за кожним з наявних атрибутів:

а) атрибут «Вік» має три значення :

« <=30 »: так – 2, ні – 3; «31..40 »: так - 4, ні - 0, « >40 »: так – 3, ні – 2.

Отримуємо:

$$\text{Entropy(Вік)} = 5/14\,(-2/5\log(2/5) - 3/5\log(3/5)) + 4/14\,(0) +$$
$$+ 5/14\,(-3/5\log(3/5) - 2/5\log(2/5)) = 5/14(0.9709) + 0 + 5/14(0.9709) = 0.6935.$$
$$\text{Gain(Вік)} = 0.94 - 0.6935 = 0.2465.$$

б) атрибут «Дохід» має три значення:

«високий»: так – 2, ні – 2; «середній»: так – 4, ні –2; «низький»: так – 3, ні – 1.

$$\text{Entropy(дохід)} = 4/14(-2/4\log(2/4)-2/4\log(2/4)) +$$
$$+ \ 6/14\ (-4/6\log(4/6) - 2/6\log(2/6)) + 4/14\ (-3/4\log(3/4)-1/4\log(1/4)) =$$
$$= 4/14\ (1) + 6/14\ (0.918) + 4/14\ (0.811) = 0.285714 + 0.393428 + 0.231714 =$$
$$= 0.9108.$$

$$\text{Gain(дохід)} = 0.94 - 0.9108 = 0.0292.$$

в) атрибут «Студент» має два значення:

«так»: так – 6, ні – 1; «ні»: так – 3, ні – 4.

$$\text{Entropy(студент)} = 7/14(-6/7\log(6/7)) + 7/14(-3/7\log(3/7)-4/7\log(4/7) =$$
$$=7/14(0.5916) + 7/14(0.9852) = 0.2958 + 0.4926 = 0.7884.$$

$$\text{Gain (Студент)} = 0.94 - 0.7884 = 0.1516.$$

в) атрибут «Кредитна історія» має два значення:

«гарна»: так – 6, ні – 2; «відмінна»: так – 3, ні – 2.

$$\text{Entropy(кредит)} = 8/14(-6/8\log(6/8)-2/8\text{lo g}(2/8)) + 6/14(-3/6\log(3/6) -$$
$$-3/6\log(3/6)) = 8/14(0.8112) + 6/14(1) = 0.4635 + 0.4285 = 0.8920$$
$$\text{Gain(Кредит)} = 0.94 - 0.8920 = 0.048.$$

Оскільки найбільше значення Gain має атрибут «Вік», починаємо побудову дерева рішень з його розбиття (рис. 3):

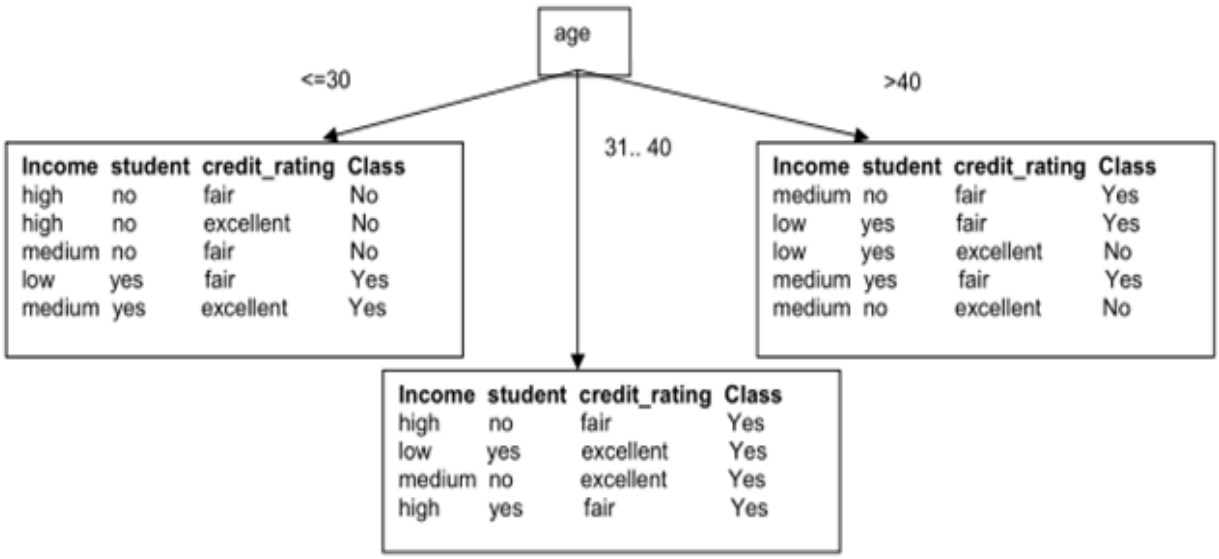

Рисунок 3 – Початок побудови дерева рішень
з розбиття атрибуту «Вік»

Оскільки всі записи у гілці з віком «31..40» належать одному класу «так», замінимо отриманий блок на лист зі значенням «Придбати» = «так» (рис. 4).

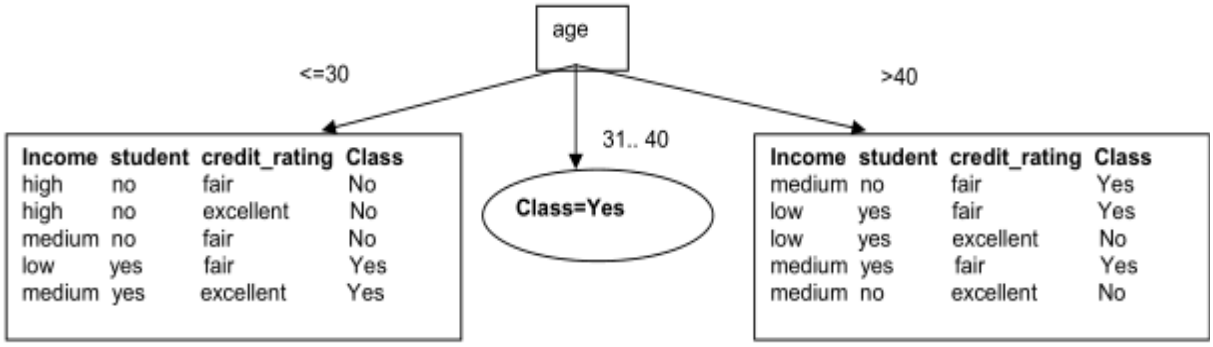

Рисунок 4 – Результат розбиття дерева рішень
по атрибуту «Вік»

3. Розглянемо, за яким атрибутом потрібно розбити гілку «Вік» = «<=30».

Ентропія блока «≤30»:

$$I(S_{ТАК}, S_{НІ}) = I(2,3) = -2/5 \log 2 (2/5) – 3/5 \log 2 (3/5) = 0.97.$$

а) атрибут «Дохід» має три значення:

«високий»: так – 0, ні – 2; «середній»: так – 1, ні – 1; «низький»: так – 1, ні – 0.

$$Entropy(Дохід) = 2/5(0) + 2/5 (-1/2\log(1/2)-1/2\log(1/2)) + 1/5 (0) = 2/5 (1) = 0,4.$$

$$Gain(Дохід) = 0,97 – 0,4 = 0,57.$$

б) атрибут «Студент» має два значення:

«так»: так – 2, ні – 0, «ні»: так – 0, ні – 3.

$$Entropy(Студент) = 2/5(0) + 3/5(0) = 0.$$

$$Gain (Студент) = 0,97 – 0 = 0,97.$$

Оскільки значення показника Gain для атрибуту «Студент» є максимальним, можна робити розбиття дерева за ним, без перевірки інших атрибутів (рис. 5).

Оскільки дві нових гілки є чистими класами, перетворимо їх на два листи з відповідними мітками (рис. 6).

Розглянемо гілку «>40», що залишилась.
Ентропія блоку «>40»:

$$I(S_{ТАК}, S_{НІ}) = I(3,2) = -3/5 \log 2 (3/5) – 2/5 \log 2 (2/5) = 0.97.$$

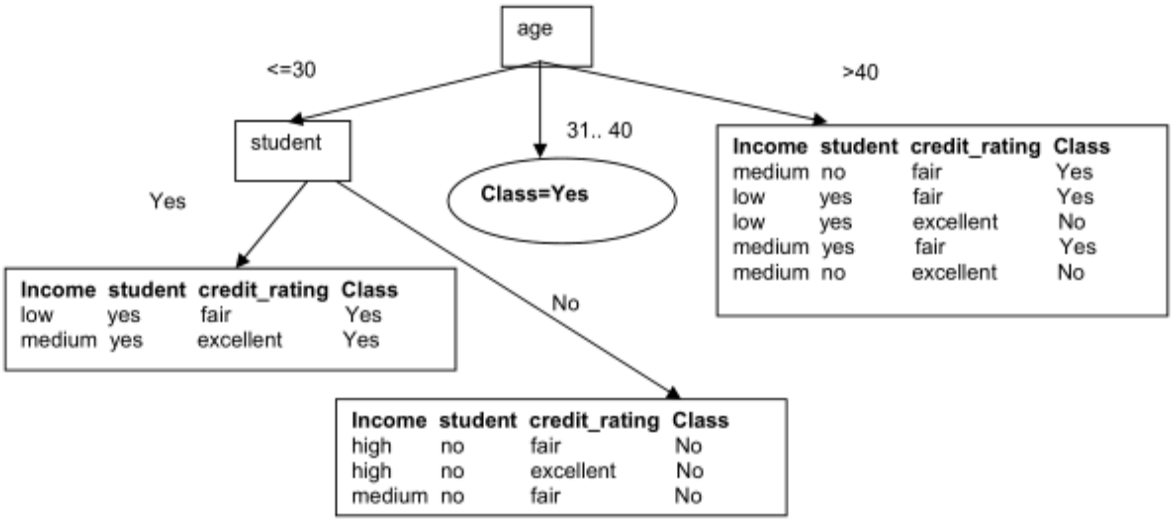

Рисунок 5 – Побудова дерева рішень розбиттям атрибуту «Студент»

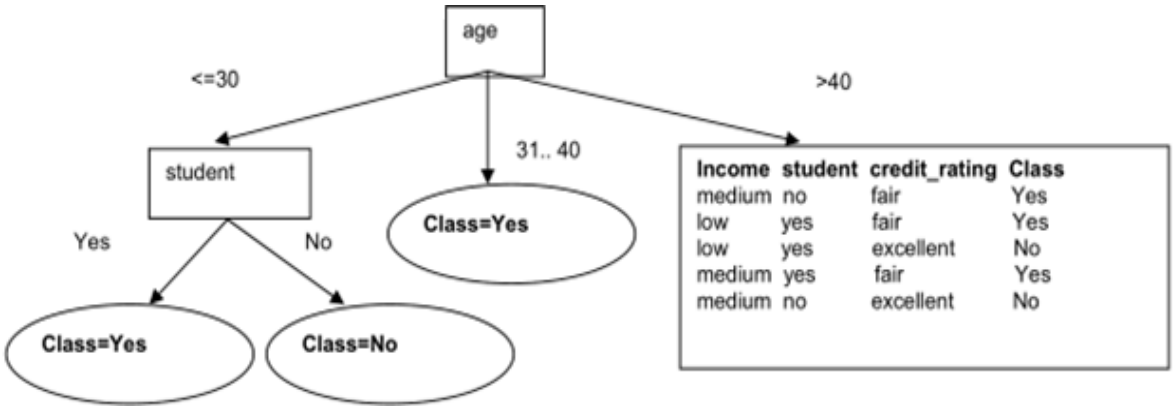

Рисунок 6 – Результат розбиття дерева рішень за атрибутом «Студент»

а) атрибут «Дохід» має два значення:

«середній»: так – 2, ні – 1; «низький»: так – 1, ні – 1.

$$Entropy(Дохід) = 3/5(-2/3\log(2/3)-1/3\log(1/3)) +$$
$$+ 2/5\ (-1/2\log(1/2) - 2\log(1/2)) = 3/5(0.9182)+2/5\ (1) = 0,55+0,4= 0,95.$$

$$Gain(Дохід) = 0,97 – 0,95 = 0,02.$$

б) атрибут «Студент» має два значення:

«так»: так – 2, ні – 1; «ні»: так – 1, ні – 1.

Entropy(Студент)=3/5(-2/3log(2/3)-1/3log(1/3))+
+2/5(-1/2log(1/2) - 1/2log(1/2))= 0,95.
Gain (Студент)=0,97–0,95 = 0,02.

в) атрибут «Кредитної історія» має два значення:

«гарна»: так – 3, ні – 0; «відмінна»: так – 0, ні – 2.

Entropy(Кредит) = 0.

Gain(Кредит) = 0,97 – 0 = 0,97.

Здійснимо розбиття за атрибутом «Кредит», яке надасть два чистих класи.

Побудоване дерево рішень набуде завершеного вигляду:

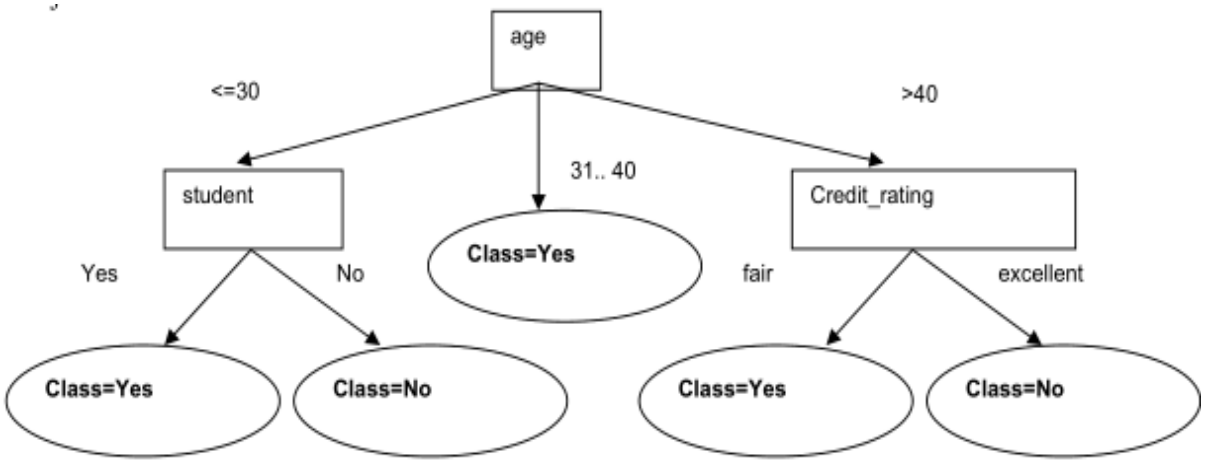

Рисунок 7 – Дерево рішень для класифікації клієнтів відносно їх ступеня зацікавленості щодо придбання комп’ютера

3.3 Кластеризація

Кластерізація здійснює угрупування об’єктів на основі близькості їх властивостей; кожен кластер складається зі схожих об’єктів, а об’єкти різних кластерів мають суттєво відрізняться [3].

3.3.1 Алгоритм розділової кластеризації k-means

Завдання: задано набір з 8 точок у двовимірному просторі (табл. 5), який треба розбити на два кластери: k=2.

Таблиця 5 – Об’єкти розділової кластеризації

A	B	C	D	E	F	G	H
(1; 3)	(3; 3)	(4; 3)	(5; 3)	(1; 2)	(4; 2)	(1; 1)	(2;1)

Розв’язання. Крок 1. Випадковим чином визначимо дві точки m1 = G і m2 = H, як центри кластерів (рис. 8).

Крок 2, ітерація 1. Для кожної точки визначимо найближчий до неї центр кластера в евклідовій метриці

$$d(X,Y) = \sqrt{\sum_i (x_i - y_i)^2},$$

тим самим визначаючи, до якого кластера вона відноситься (табл. 6).

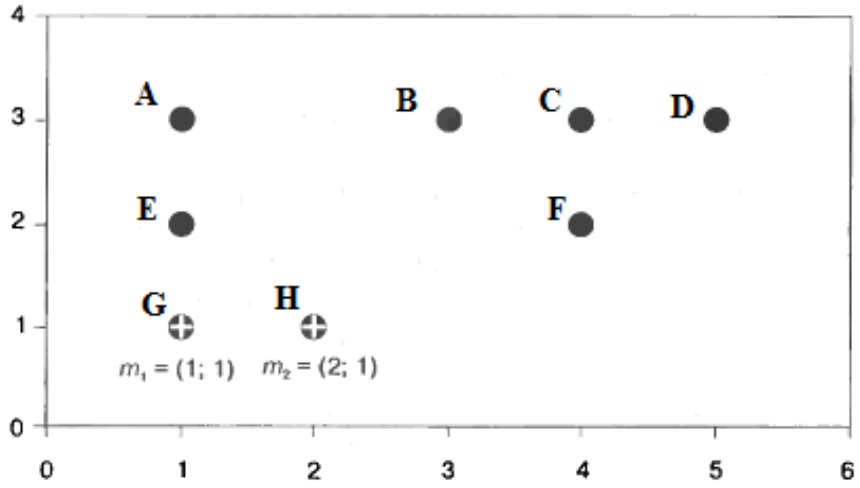

Рисунок 8 – Початкова ініціалізація задачі кластеризації

Таблиця 6 – Визначення для кожного об’єкта кластеризації найближчого з центрів кластера (перша ітерація)

Об’єкт	A	B	C	D	E	F	G	H
Відстань до m1	2,00	2,83	3,61	4,47	1,00	3,16	0,00	1,00
Відстань до m2	2,24	2,24	2,83	3,61	1,41	2,24	1,00	0,00
Кластер	1	2	2	2	1	2	1	2

Крок 3, ітерація 1. Обчислимо центроїди, до яких переміщаються центри кластерів (рис. 9).

Ц1= [(1+1+1/3); (3+2+1/3)]=(1; 2);

Ц2=[(3+4+5+4+2/5); (3+3+3+2+1/5)]=(3,6; 2,4).

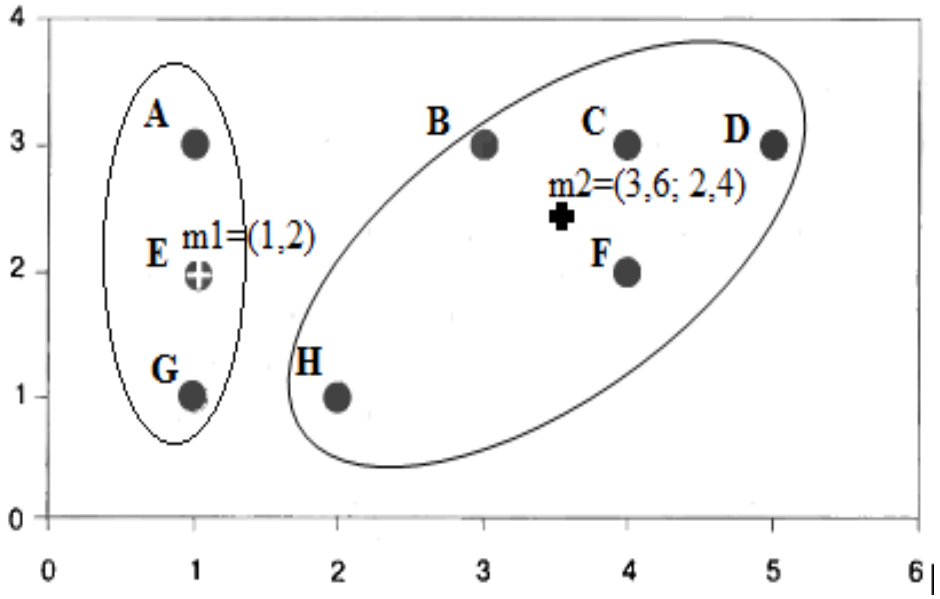

Рисунок 9 – Кластери і центроїди, визначені після першої ітерації алгоритму

Визначивши належність точок кластерам, обчислюємо суму квадратів помилок:

$$E = \sum_{i=1}^{k} \sum_{p \in C} (p - m)^2 = 2^2 + 2{,}24^2 + 2{,}83^2 + 3{,}61^2 + 1^2 + 2{,}24^2 + 0^2 + 0^2 = 36.$$

Крок 2, ітерація 2. Для кожного об'єкта кластеризації знов обчислимо найближчий до нього центр нових кластерів та визначимо належність об'єкта найближчому кластеру (табл. 7).

Таблиця 7 – Визначення для кожного об'єкта кластеризації найближчого з центрів кластера (друга ітерація)

Об'єкт	A	B	C	D	E	F	G	H
Відстань до m1	1,00	2,24	3,16	4,12	0,00	3,00	1,00	1,41
Відстань до m2	2,67	0,85	0,72	1,52	2,63	0,57	2,95	2,13
Кластер	1	2	2	2	1	2	1	1

Відносно велика зміна місця розташування центра другого кластера m2 призвела до того, що місце розташування точки H стало ближчим до центра m1, отже вона увійшла до складу першого кластера. Нова сума квадратів помилок становила:

$$E = \sum_{i=1}^{k} \sum_{p \in C} (p - m)^2 = 1^2 + 0{,}85^2 + 0{,}72^2 + 1{,}52^2 + 0^2 + 0{,}57^2 + 1^2 + 1{,}41^2 = 7{,}86.$$

Помилка зменшилась, що означає краще групування об’єктів відносно центрів кластерів.

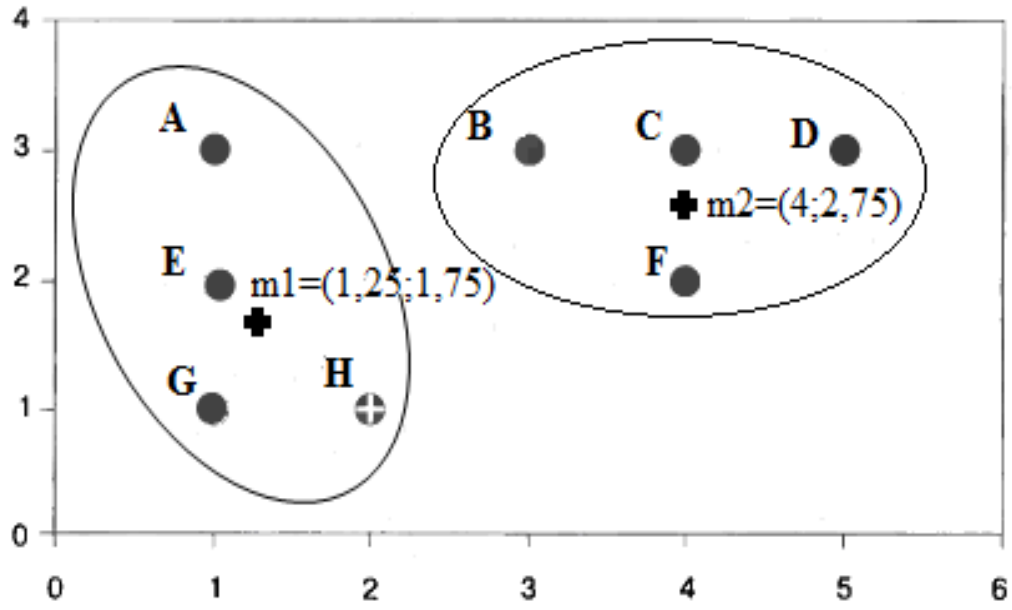

Рисунок 10 – Кластери і центроїди, визначені після першої ітерації алгоритму

Крок 3, ітерація 2. Обчислимо нові центроїди для кожного кластера:

Ц1=[(1+1+1+2/4);(3+2+1+1/4)]=(1,25;1,75);

Ц2= [(3+4+5+4/4);(3+3+3+2+/4)]=(4;2,75).

Порівняно з минулим проходом центри кластерів змінилися не суттєво. Мало змінилася відносно першої ітерації і сума квадратів помилок.

Крок 2, ітерація 3. Визначимо відстані об’єктів від найближчого з центрів нових кластерів (табл. 8).

Таблиця 8 – Визначення для кожного об’єкта кластеризації найближчого з центрів кластера (третя ітерація)

Об’єкт	A	B	C	D	E	F	G	H
Відстань до m1	1,27	2,15	3,92	3,95	0,35	2,76	0,79	1,06
Відстань до m2	3,01	1,03	0,25	1,03	3,09	0,75	3,47	2,66
Кластер	1	2	2	2	1	2	1	1

Нова сума квадратів помилок:

$$E=\sum_{i=1}^{k}\sum_{p\in C}(p-\ m)^2=$$
$$=1{,}27^2+1{,}03^2+0{,}25^2+1{,}03^2+0{,}35^2+0{,}57^2+0{,}79^2+1{,}06^2=6{,}23.$$

Сума квадратів помилок мала змінилась відносно попереднього проходу.

Крок 3, прохід 3. Обчислюємо нові центроїди кластерів. Оскільки жодний об’єкт не змінив свого членства у кластерах і положення центроїдів практично не змінилося, алгоритм завершує свою роботу.

3.3.2 Алгоритм ієрархічної агломеративної кластеризації

Завдання: задано набір з 5 об’єктів кластеризації в одновимірному просторі (табл. 9). Використайте алгоритм агломеративної ієрархічної кластеризації з евклідовою мірою відстані і критерієм максимальної відстані.

Побудуйте дендограму виявлених кластерів. Скільки кластерів буде виділено при використанні як поріг обрізання максимального часу життя кластерів? Скільки об’єктів буде входити до кожного кластера?

Таблиця 9 – Об’єкти ієрархічної кластеризації

A	B	C	D	E
1	5	8	9	2

Розв’язання. Для здійснення кластеризації побудуємо матрицю відстаней об’єктів D_0:

$$D_0 = \begin{array}{c|ccccc} & \mathbf{A} & \mathbf{B} & \mathbf{C} & \mathbf{E} & \mathbf{F} \\ \hline \mathbf{A} & 0 & 4 & 7 & 9 & 1 \\ \mathbf{B} & 4 & 0 & 3 & 5 & 3 \\ \mathbf{C} & 7 & 3 & 0 & 2 & 6 \\ \mathbf{E} & 9 & 5 & 2 & 0 & 8 \\ \mathbf{F} & 1 & 3 & 6 & 8 & 0 \end{array}$$

Оскільки мінімальною у матриці є відстань між об’єктами A і F, об’єднуємо їх в один кластер з часом життя 1.

Згідно з застосуванням критерію максимальної відстані визначимо відстань між кластером {a,f} та іншими об’єктами, кожен з яких на цей момент подає один окремий кластер:

d (b, {a,f}) = max {d (b,a), d (b,f)} = max {4, 3} = 4;
d (c, { a,f }) = max {d (c,a), d (c,f)} = max {7, 6} = 7;
d (e, { a,f }) = max {d (e,a), d (e,f)} = max {9, 8} = 9.

Побудуємо оновлену матрицю відстаней D_1:

$D_1=$

	{a,f}	**B**	**C**	**E**
{a,f}	0	4	7	9
B	4	0	3	5
C	7	3	0	2
E	9	5	2	0

Бачимо, що найменшою є відстань між об’єктами C і E. Отже, створимо з них новий кластер {c,e}.

Для використання критерію максимальної відстані знов визначимо відстані між кластерами {a,f}, B та {c,e}:

d ({a,f}, {c, e}) = max {d ({a,f}, c), d ({a,f}, e)} = max {7, 9} = 9;
d (b, {c,e}) = max {d (b,c), d (b,e)} = max {3,5} = 5.

Побудуємо оновлену матрицю відстаней D_2:

$D_2=$

	{a,f}	**B**	**{c,e}**
{a,f}	0	4	7
B	4	0	5
{c,e}	7	5	0

Виконуючу втретє ту саму процедуру, об’єднуємо об’єкт B з кластером {a,f}, внаслідок чого отримуємо новий кластер {a,b,f} з матрицею відстаней D_3:

$D_3=$

	{a,b,f}	**{c,e}**
{a,b,f}	0	5
{c,e}	5	0

Побудуємо дендограму, яка містить інформацію про всі кластери, що були створені в процесі кластеризації вихідної множини об’єктів (рис. 11).

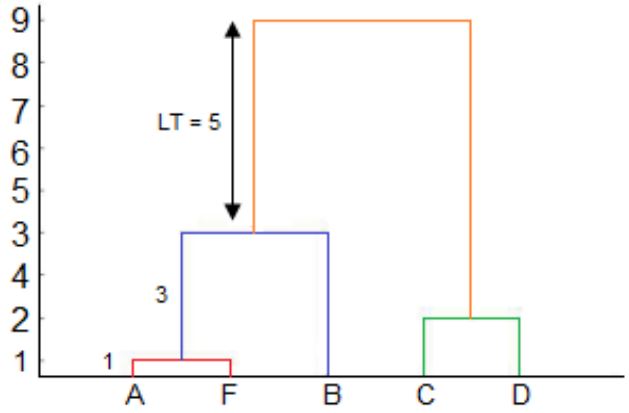

Рисунок 11 – Дендограма агломеративної кластеризації

Найбільший показник часу життя (Life time) знаходиться на третьому етапі різання дендограми, на якому породжує два кластери: {a, b, f} і {c, e}. Всього у процесі кластеризаціє було побудовано 4 кластери (рис. 12).

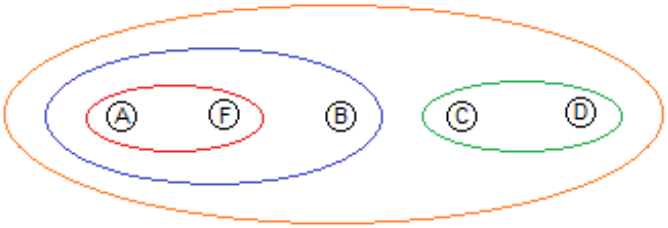

Рисунок 12 – Кластери, отримані під час ієрархічної аглометричної кластеризації

3.3.3 Кластеризація з використанням мережі Кохонена

Завдання. Для кластеризації клієнтів банку за двома параметрами: віком та доходами, кожен з яких має по два значення. Потрібно здійснити навчання мережі Кохонена з використанням наданої банком навчальної вибірки (табл. 10) [3].

Таблиця 10 – Навчальна вибірка для кластеризації

№	x_{i1}	x_{i2}	Опис
1	x_{11}=0,8	x_{12}=0,8	Літня людина з високим доходом
2	x_{21}=0,8	x_{22}=0,1	Літня людина з низьким доходом
3	x_{31}=0,2	x_{32}=0,8	Молода людина з високим доходом
4	x_{41}=0,1	x_{42}=0,2	Молода людина з низьким доходом

Розв'язання. Вихідний шар мережі Кохонена буде мати 4 нейрони (рис. 12), оскільки їх кількість має дорівнювати кількості створюваних кластерів.

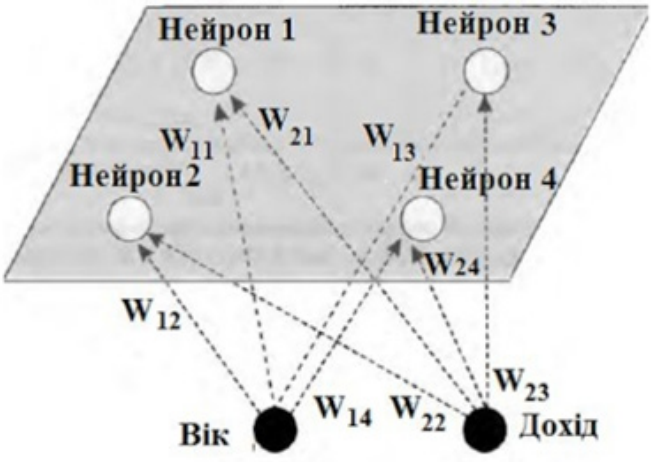

Рисунок 12 – Структура мережі Кохонена для кластеризації клієнтів за двома значеннями кожного з параметрів: «Вік» і «Доход»

На етапі ініціалізації, через надто малий розмір мережі, встановимо радіус навчання R=0. Таким чином, можливість налаштування вагових коефіцієнтів буде надано лише нейрону–переможцю. Коефіцієнт швидкості навчання встановимо на рівень η = 0,5 і згенеруємо випадковим чином початкові ваги нейронів (табл. 11).

Таблиця 11 – Початкові ваги нейронів

w_{11}	w_{21}	w_{12}	w_{22}	w_{13}	w_{23}	w_{14}	w_{24}
0,9	0,8	0,9	0,2	0,1	0,8	0,1	0,2

Ітерація 1. На етапі збудження подамо на перший нейрон перший вектор впливу з навчальної вибірки X_1 (0,8; 0,8).

На етапі конкуренції обчислимо евклідову відстань між вхідним вектором X_1 і векторами ваг усіх чотирьох нейронів вихідного шару:

Нейрон 1: $D(W_1, X_1) = \sqrt{(w_{11} - x_{11})^2 + (w_{21} - x_{12})^2} = \sqrt{(0{,}9 - 0{,}8)^2 + (0{,}8 - 0{,}8)^2} = 0{,}1.$

Нейрон 2: $D(W_2, X_1) = \sqrt{(w_{12} - x_{11})^2 + (w_{22} - x_{12})^2} = \sqrt{(0{,}9 - 0{,}8)^2 + (0{,}2 - 0{,}8)^2} = 0{,}61.$

Нейрон 3: $D(W_3, X_1) = \sqrt{(w_{13} - x_{11})^2 + (w_{23} - x_{12})^2} = \sqrt{(0{,}1 - 0{,}8)^2 + (0{,}8 - 0{,}8)^2} = 0{,}7.$

Нейрон 4: $D(W_4, X_1) = \sqrt{(w_{14} - x_{11})^2 + (w_{24} - x_{12})^2}$
$= \sqrt{(0{,}1 - 0{,}8)^2 + (0{,}2 - 0{,}8)^2} = 0{,}92.$

Найменшу відстань між вектором власних ваг W_1 = (0,9; 0,8) і вхідним вектором навчальної вибірки X_1 = (0,8; 0,8) має нейрон 1. Отож, саме він перемагає в конкуренції за право відкрити кластер, куди будуть вноситися літні люди з високим доходом.

Оскільки через малий розмір мережі радіус навчання має значення R=0, етап об'єднання, що мав би містити сусідні нейрони переможця, пропускається.

На етапі налаштування здійснимо налаштування ваг лише нейрона 1 (нейрона-переможця) згідно з формулою:

$$w_{ij,\text{нове}} = w_{ij,\text{поточне}} + \eta \cdot (x_{ni} - w_{ij,\text{поточне}}),$$

де η = 0,5;

j=1 для першого нейрона;

n=1 для першого запису.

Отже, для ознаки «Вік» отримуємо:

$$w_{11,\text{нове}} = w_{11,\text{поточне}} + 0{,}5 \cdot (x_{11} - w_{11,\text{поточне}}) = 0{,}9 + 0{,}5 \cdot (0{,}8 - 0{,}9) = 0{,}85.$$

Для ознаки «Дохід» отримуємо:

$$w_{21,\text{нове}} = w_{21,\text{поточне}} + 0{,}5 \cdot (x_{12} - w_{21,\text{поточне}}) = 0{,}8 + 0{,}5 \cdot (0{,}8 - 0{,}8) = 0{,}8.$$

Ітерація 2. Застосуємо дії, аналогічні діям ітерації 1, до другого вектора навчальної вибірки X_2 = (0,8; 0,1).

Нейрон 1: $D(W_1, X_2) = \sqrt{(w_{11} - x_{21})^2 + (w_{21} - x_{22})^2} =$
$= \sqrt{(0{,}85 - 0{,}8)^2 + (0{,}8 - 0{,}1)^2} = 0{,}71.$

Нейрон 2: $D(W_2, X_2) = \sqrt{(w_{12} - x_{21})^2 + (w_{22} - x_{22})^2} =$
$= \sqrt{(0{,}9 - 0{,}8)^2 + (0{,}2 - 0{,}1)^2} = 0{,}14.$

Нейрон 3: $D(W_3, X_2) = \sqrt{(w_{13} - x_{21})^2 + (w_{23} - x_{22})^2} =$
$= \sqrt{(0{,}1 - 0{,}8)^2 + (0{,}8 - 0{,}1)^2} = 0{,}99.$

Нейрон 4: $D(W_4, X_2) = \sqrt{(w_{14} - x_{21})^2 + (w_{24} - x_{22})^2}$
$= \sqrt{(0{,}1 - 0{,}8)^2 + (0{,}2 - 0{,}1)^2} = 0{,}71.$

Найближчим до вхідного вектора другого запису з навчальної вибірки X_2= (0,8; 0,1), порівняно з векторами ваг інших нейронів, виявився вектор ваг другого нейрона W_2=(0,9; 0,2). Отже, саме він і буде «засновником» другого кластера, що міститиме записи про літніх людей з низькими доходами.

Для нейрона 2 і другого запису навчальної вибірки мають місце значення j = 2 та n = 2, відповідно. Отже, налаштування вагових коефіцієнтів другого нейрона відбудеться за формулою:

$$w_{i2,\text{нове}} = w_{i2,\text{поточне}} + \eta \cdot (x_{2i} - w_{i2,\text{поточне}}).$$

Для ознаки «Вік» оновлене значення ваги становитиме:

$$w_{12,\text{нове}} = w_{12,\text{поточне}} + 0{,}5 \cdot (x_{21} - w_{12,\text{поточне}}) = 0{,}9 + 0{,}5 \cdot (0{,}8 - 0{,}9) = 0{,}85.$$

Ознака «Дохід» отримає таке оновлене значення ваги:

$$w_{22,\text{нове}} = w_{22,\text{поточне}} + 0{,}5 \cdot (x_{22} - w_{22,\text{поточне}}) = 0{,}2 + 0{,}5 \cdot (0{,}1 - 0{,}2) = 0{,}15.$$

Ітерація 3. Виконаємо аналогічні дії з третім вектором навчальної вибірки X_3 = (0,2; 0,9).

Нейрон 1: $$D(W_1, X_3) = \sqrt{(w_{11} - x_{31})^2 + (w_{21} - x_{32})^2} = \sqrt{(0{,}85 - 0{,}2)^2 + (0{,}8 - 0{,}9)^2} = 0{,}66.$$

Нейрон 2: $$D(W_2, X_3) = \sqrt{(w_{12} - x_{31})^2 + (w_{22} - x_{32})^2} = \sqrt{(0{,}85 - 0{,}2)^2 + (0{,}15 - 0{,}9)^2} = 0{,}99.$$

Нейрон 3: $$D(W_3, X_3) = \sqrt{(w_{13} - x_{31})^2 + (w_{23} - x_{32})^2} = \sqrt{(0{,}1 - 0{,}2)^2 + (0{,}8 - 0{,}9)^2} = 0{,}14.$$

Нейрон 4: $$D(W_4, X_3) = \sqrt{(w_{14} - x_{31})^2 + (w_{24} - x_{32})^2} = \sqrt{(0{,}1 - 0{,}2)^2 + (0{,}2 - 0{,}9)^2} = 0{,}71.$$

Найближчим до вхідного вектора третього запису з навчальної вибірки X_3= (0,2; 0,9) виявився вектор ваг третього нейрона W_3=(0,1; 0,8). Таким чином, він стане «засновником» кластера для молодих людей з високими доходами.

Для нейрона 3 і третього запису навчальної вибірки мають місце значення j = 3 та n = 3, відповідно. Отже, налаштування вагових коефіцієнтів третього нейрона відбудеться за формулою:

$$w_{i3,\text{нове}} = w_{i3,\text{поточне}} + \eta \cdot (x_{3i} - w_{i3,\text{поточне}}),$$

Для ознаки «Вік» оновлене значення ваги становитиме:

$$w_{13,\text{нове}} = w_{13,\text{поточне}} + 0{,}5 \cdot (x_{31} - w_{13,\text{поточне}}) = 0{,}1 + 0{,}5 \cdot (0{,}2 - 0{,}1) = 0{,}15.$$

Ознака «Дохід» отримає таке оновлене значення ваги:

$$w_{23,\text{нове}} = w_{23,\text{поточне}} + 0{,}5 \cdot (x_{32} - w_{23,\text{поточне}}) = 0{,}8 + 0{,}5 \cdot (0{,}9 - 0{,}8) = 0{,}85.$$

Ітерація 4. Здійснимо навчання мережі Кохонена на четвертому записі з навчальної вибірки $X_{(4)} = (0{,}1; 0{,}1)$.

Нейрон 1: $D(W_1, X_4) = \sqrt{(w_{11} - x_{41})^2 + (w_{21} - x_{42})^2} = \sqrt{(0{,}85 - 0{,}1)^2 + (0{,}8 - 0{,}1)^2} = 1{,}05.$

Нейрон 2: $D(W_2, X_4) = \sqrt{(w_{12} - x_{41})^2 + (w_{22} - x_{42})^2} = \sqrt{(0{,}85 - 0{,}1)^2 + (0{,}15 - 0{,}1)^2} = 0{,}75$

Нейрон 3: $D(W_3, X_4) = \sqrt{(w_{13} - x_{41})^2 + (w_{23} - x_{42})^2} = \sqrt{(0{,}1 - 0{,}1)^2 + (0{,}8 - 0{,}1)^2} = 0{,}7.$

Нейрон 4: $D(W_4, X_4) = \sqrt{(w_{14} - x_{41})^2 + (w_{24} - x_{42})^2} = \sqrt{(0{,}1 - 0{,}1)^2 + (0{,}2 - 0{,}1)^2} = 0{,}1.$

Найближчим до вхідного вектора четвертого запису з навчальної вибірки $X_4 = (0{,}1;\ 0{,}1)$ виявився вектор ваг четвертого нейрона $W_4 = (0{,}1;\ 0{,}2)$. Таким чином, він стає «засновником» кластера для молодих людей з низьким доходом.

Для нейрона 4 і четвертого запису навчальної вибірки мають місце значення $j = 4$ та $n = 4$, відповідно. Отже, налаштування вагових коефіцієнтів четвертого нейрона відбудеться за формулою:

$$w_{i4,\text{нове}} = w_{i4,\text{поточне}} + \eta \cdot (x_{4i} - w_{i4,\text{поточне}}),$$

Для ознаки «Вік» оновлене значення ваги не зміниться:

$$w_{14,\text{нове}} = w_{14,\text{поточне}} + 0{,}5 \cdot (x_{41} - w_{14,\text{поточне}}) = 0{,}1 + 0{,}5 \cdot (0{,}1 - 0{,}1) = 0{,}1.$$

Ознака «Дохід» отримає оновлене значення ваги:

$$w_{24,\text{нове}} = w_{24,\text{поточне}} + 0{,}5 \cdot (x_{42} - w_{24,\text{поточне}}) = 0{,}2 + 0{,}5 \cdot (0{,}1 - 0{,}2) = 0{,}15.$$

Отже, під час навчання мережі Кохонена чотири вихідні нейрони налаштовані на формування 4-х окремих кластерів, наведених у таблиці 12.

Таблиця 12 – Результати кластеризації

№ кластера	№ нейрона	Опис кластера
1	1	Літня людина з високим доходом
2	2	Літня людина з низьким доходом
3	3	Молода людина з високим доходом
4	4	Молода людина з низьким доходом

3.4 Асоціативні правила

Методи побудови асоціативних правил призначені для пошуку зв'язків між подіями, що відбуваються сумісно, і виявлення правил, (асоціацій) для їх кількісного опису. Множина подій, що відбуваються сумісно, називається транзакцією [3].

Завдання.

База даних містить шість транзакцій (табл. 13). Знайдіть асоціативні правила при рівні мінімальної підтримки S = 50% і мінімальній точності (достовірності) C = 80%.

Таблиця 13 – Набір транзакцій

ID	Транзакція
T1	A,B,C
T2	B,D
T3	B,A,D,C
T4	E,D
T5	A,B,C,D
T6	F

Розв'язання.

1. Визначимо (табл. 14), які з об'єктів (одноелементних наборів) A…F, задовольняють заданий в умовах завдання рівень підтримки C (частоту появи в наведеному наборі транзакцій):

$$\text{Загальна C } (A \rightarrow B) = P(A \cap B)/P(A) = \frac{\text{Кількісь транзакцій,що містять A і B}}{\text{Загальна кількість транзакцій}}$$

Таблиця 14 – Рівень підтримки окремих об'єктів у наборі транзакцій

Об'єкт	Кількість появ у наборі транзакцій	Рівень підтримки S
A	3	3/6 = 50%
B	4	4/6 = 75%
C	3	3/6 = 50%
D	4	4/6 = 75%
E	1	1/6 = 17%
F	1	1/6 = 17%

2. З об’єктів A…D, що задовольняють вимогу до рівня підтримки S ≥ 50%, сформуємо двохелементні набори і визначимо рівень підтримки цих наборів (табл. 15).

Таблиця 15 – Рівень підтримки двохелементних наборів об’єктів у наборі транзакцій

Двохелементні набори об’єктів	Кількість появ у наборі транзакцій	Рівень підтримки S
AB	3	3/6 = 50%
AC	3	3/6 = 50%
AD	2	2/6 = 33%
BC	2	2/6 = 33%
BD	3	3/6 = 50%
CD	2	1/6 = 33%

3. З двохелементних наборів об’єктів AB, AC, BD, що задовольняють вимогу до рівня підтримки S ≥ 0,5, сформуємо трьохелементні набори і визначимо рівень підтримки цих наборів (табл. 16).

Таблиця 16 – Рівень підтримки трьохелементних наборів об’єктів у наборі транзакцій

Трьохелементні набори об’єктів	Кількість появ у наборі транзакцій	Рівень підтримки S
ABC	3	3/6 = 50%
ABD	2	3/6 = 33%
ACD	2	2/6 = 33%
BCD	2	2/6 = 33%

4. З виявлених наборів об’єктів з підтримкою S ≥ 50% сформуємо всі можливі варіанти асоціативних залежностей (правил) виду «якщо A, то B» і обчислимо рівень їх достовірності (надійності) C (confidence) за формулою:

$$C\ (A \rightarrow B) = P(B/A);$$

$$P(A \cap B) = P(A) \times P(B/A);$$

$$C = P(A \cap B)/P(A) = \frac{\text{Кількісь транзакцій,що містять A і B}}{\text{Кількісь транзакцій,що містять A без B}}$$

Зведемо отримані результати до таблиці 17.

Таблиця 17 – Асоціативні правила, виявлені в заданому наборі транзакцій

Асоціативні правила	Підтримка	Достовірність
A→B	3/6 = 50%	3/3 = 100%
B→A	3/6 = 50%	3/4 = 75%
A→C	3/6 = 50%	3/3 = 100%
C→A	3/6 = 50%	3/3 = 100%
B→D	3/6 = 50%	3/4 = 75%
D→B	3/6 = 50%	3/4 = 75%
A→BC	3/6 = 50%	3/3 = 100%
B→CA	3/6 = 50%	3/4 = 75%
C→AB	3/6 = 50%	3/3 = 100%
AB→C	3/6 = 50%	3/3 = 100%
BC→A	3/6 = 50%	3/3 = 100%
CA→B	3/6 = 50%	3/3 = 100%

У таблиці 18 наведено асоціативні правила, що відповідають заданим обмеженням з підтримки та точності.

Таблиця 18 – Асоціативні правила, що відповідають заданим обмеженням з підтримки та точності

Асоціативні правила	Підтримка	Достовірність
A→B	3/6 = 50%	3/3 = 100%
A→C	3/6 = 50%	3/3 = 100%
C→A	3/6 = 50%	3/3 = 100%
A→BC	3/6 = 50%	3/3 = 100%
C→AB	3/6 = 50%	3/3 = 100%
AB→C	3/6 = 50%	3/3 = 100%
BC→A	3/6 = 50%	3/3 = 100%
CA→B	3/6 = 50%	3/3 = 100%

3.5 Генетичний алгоритм

Генетичний алгоритм – це еволюційний алгоритм пошуку, що використовується для вирішення задач оптимізації і моделювання шляхом послідовного підбору, комбінування та варіації шуканих параметрів з використанням механізмів, що нагадують біологічну еволюцію.

Завдання. Продемонструвати одне покоління (ітерацію) генетичного алгоритму при розв’язанні діофантова рівняння: a +3b + 5c = 12.

Розв'язання. Діофантовим називається багаточлен з цілочисельними коефіцієнтами і цілочисельними розв'язками.

1. Визначимо форму подання хромосоми, яка подає можливі розв'язки задачі. Розв'язком в такому випадку є значення змінних *a, b* і *c* , а хромосома генетичного алгоритму має мати бінарне подання. Отже, подамо хромосому послідовністю двійкових кодів значень *a, b* і *c* (рис. 13).

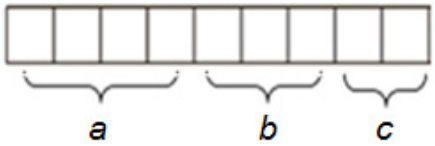

Рисунок 13 – Структура хромосоми
для розв'язання поставленої задачі

Визначимо діапазони значень, що можуть набувати змінні *a, b* і *c*:

$$0 \le a, b \le 4; 0 \le c \le 2.$$

Приклади можливих хромосом і двійковій та десятковій формах наведені на рисунку 14.

2. Створимо з наведених на рисунку 14 випадково згенерованих хромосом (можливих розв'язків задачі) початкову популяцію генетичного алгоритму.

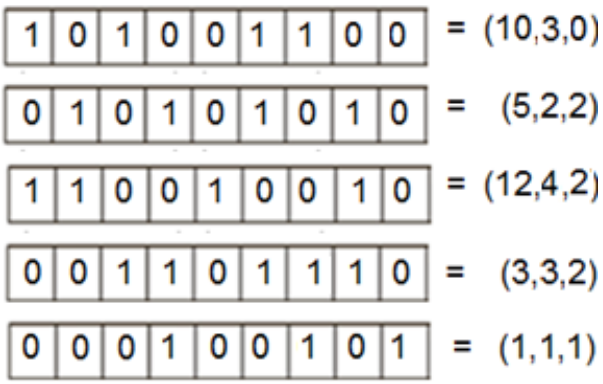

Рисунок 14 – Хромосоми початкової популяції
генетичного алгоритму

3. Визначимо (табл. 19) значення функції пристосованості (фітнес функції) кожної з хромосом до середовища задачі (у нашому випадку ступеня відповідності випадкового розв'язку правильному).

Таблиця 19 – Обчислення значень функцій пристосованості хромосом початкової популяції

№	Хромосома	$F_{\text{пристасованості}}$
1	**10** + 3×**3** + 5×**0** = 24	|19 – 12| = 7
2	**5** + 3×**2** + 5×**2** = 21	|21 – 12| = 9
3	**12** + 3×**4** + 5×**2** = 40	|34 – 12| = 22
4	**3** + 3×**3** + 5×**2** = 25	|22 – 12| = 10
5	**1** + 3×**1** + 5×**1** = 18	| 9 – 12| = 3
	Середнє значення:	10,2

4. Відбір хромосом для створення батьківських пар здійснимо методом рулетки, відповідно до якого ширина сектора рулетки, що визначає відбір хромосоми до батьківській пари, є пропорційною значенню її функції пристосованості. Оскільки кращими є менші значення функції пристосованості, то імовірність відбору хромосоми до батьківської пари буде обернено пропорційною значенню її функції пристосованості (табл. 18).

Будемо породжувати в кожній популяції 6 нових хромосом, для чого знадобиться 6 батьків. Отже, кожні 17% ширини сектора рулетки, нададуть відповідній хромосомі право одного входження до батьківської пари. Згідно з розрахунками (табл. 20) хромосоми 2 (імовірність 28%) і 4 (імовірність 25%) увійдуть до складу трьох батьківських пар, хромосома 5 (20%) – до двох, і хромосоми 1 і 3 – до однієї батьківської пари.

5. Випадковим чином сформуємо батьківські пари і випадковим чином призначимо кожній парі один з трьох можливих варіантів схрещування: обмін генами фрагментів *a*, *b* або *c* батьківських хромосом (рис. 15). Для спрощення сприйняття будемо записувати хромосоми в десятковій системі числення.

Таблиця 20 – Обчислення імовірності відбору хромосом початкової популяції до батьківських пар

№	Обернене значення функції пристосованості	Імовірність	Батьківство
1	$7^{-1} = 0{,}143$	0,143/0,733 = 19,5%	1
2	$9^{-1} = 0{,}111$	0,111/0,733 = 15,2%	1
3	$22^{-1} = 0{,}046$	0,046/0,733 = 6,3%	0
4	$10^{-1} = 0{,}100$	0,100/0,733 = 13,6%	1
5	$3^{-1} = 0{,}333$	0,333/0,733 = 45,4%	3
	$\sum = 0{,}733$		

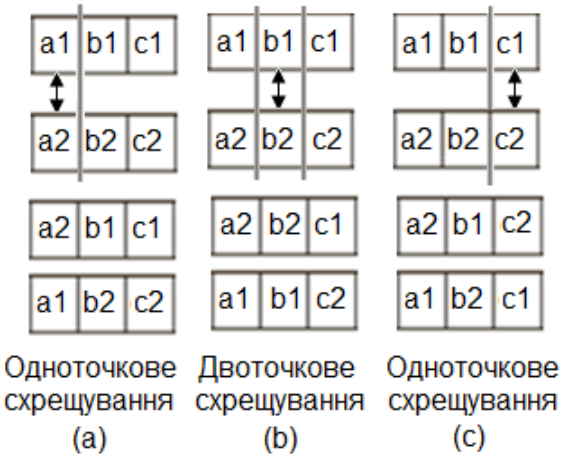

Рисунок 15 – Одноточкові (*a*, *c*) і двоточкове схрещування (*b*)

Результати схрещування відображено в таблиці 19.

6. Задамо імовірність мутації 5%. Отже, випадково відібрані 4% з 54 бітів, що містять шість дочірніх хромосом, мають змінити своє значення на протилежне. Припустимо, значення змінили 8-й біт нащадка (10,1,0) та 3-й біт нащадка (1,1,2), як це показано на рис. 16. Мутовані хромосоми та їх функції пристосованості наведено в таблиці 21.

Рисунок 16 – Мутація двох хромосом-нащадків

7. Сформуємо нову популяцію, відібравши 5 кращих з 11 наявних (5 батьківських та 6 дочірніх) хромосом за значеннями їх функцій пристосованості (табл. 22).

Як бачимо, середня пристосованість хромосом другого покоління популяції зросла майже в 4 рази – з 10,2 (див. табл. 19) до 2,7 (табл.22).

Таблиця 21 – Результати виконання генетичних операцій схрещування та мутації

№	Схрещування	Батько 1	Батько 2	Нащадки	Мутація	$\boldsymbol{F}_{\text{прист}}$
1	*b*	(1,1,1)	(10,3,0)	(1,3,1)		|15 – 12| = 3
2				(**10**,1,0)	(**14**,1,0)	|17 – 12| = 5
3	*a*	(1,1,1)	(5,2,2)	(5,1,1)		|13 – 12| = 1
4				(1,2,2)		|17 – 12| = 5
5	*c*	(1,1,1)	(3,3,2)	(1,**1**,2)	(1,**0**,2)	|11 – 12| = 1
6				(3,3,1)		|17 – 12| = 5

Таблиця 22 – Хромосоми другого покоління популяції

№	Хромосома	$F_{\text{пристосованості}}$
1	(5,1,1)	1
2	(1,0,2)	1
3	(1,1,1)	3
4	(1,3,1)	3
5	(3,3,1) або (1,2,2) або (14,1,0)	5
	Середнє значення:	2,7

Зауважимо, що в цьому невеличкому прикладі, де всі початкові значення були згенеровані випадково, генетичний алгоритм був здатний знайти правильний розв'язок задачі вже у першому поколінні. Це могло б статися, якщо б для хромосом (1,1,1) і (5,2,2) замість схрещування типу *a* було б вибрано тип схрещування *b* або для схрещування типу *c* було б вибрано хромосоми (10,3,0) і (3,3,2), (рис. 17).

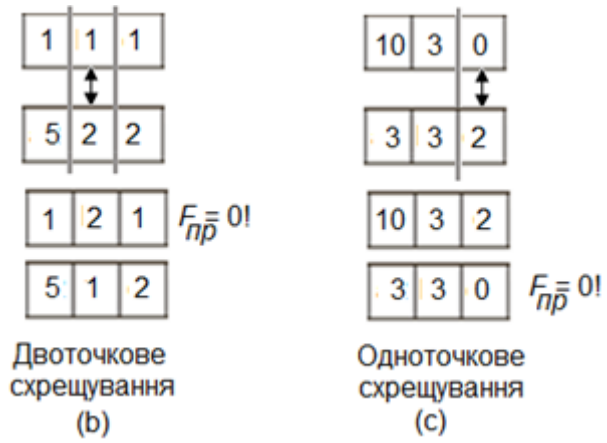

Рисунок 17 – Отримання розв'язку генетичним алгоритмом

4 ПЕРЕЛІК ПИТАНЬ ДО ІСПИТУ

1. Поняття терміна «аналіз даних».
2. Порівняння класичного підходу до аналізу даних у обчислювальній математиці та при ідентифікації.
3. Поняття ідентифікації моделі
4. Концепція та основні етапи аналізу даних за Дж. Тьюкі.
5. Поняття моделі та основні властивості моделі.
6. Поняття даних. Основні типи даних. Принципи формалізації даних.
7. Порівняльний аналіз понять даних, інформації та знань. Наведіть власні приклади кожного поняття.
8. Поняття структуризації даних. Основні властивості структурованих даних.
9. Основні типи вимірювальних шкал та їх суть.
10. Поняття бізнес-аналітики. Особливості задач бізнес-аналітики.
11. Поняття видобування знань з баз даних (Knowledge Discovery in Databases-KDD). Основні стадії KDD.
12. Поняття Data Mining. Основні задачі Data Mining.
13. Методологія CRIPS DM. Основні етапи методології.
14. Основні принципи збору даних.
15. Основні методи збору даних.
16. Системи оперативного аналізу даних OLTP. Причини недостатності OLTP для задач бізнес-аналізу.
17. Системи підтримки прийняття рішень (СППР). Відмінність організації даних у СППР та у OLTP.
18. Інтеграція даних.
19. Консолідація даних.
20. Федералізація даних.
21. Корпоративні дані. Типи корпоративних даних.
22. Оперативний склад даних. Основні переваги його використання.
23. Зона тимчасового зберігання та її основні функції.
24. Процеси ELT та ETL.
25. Поняття якості даних. Найбільш критичні фактори якості даних.
26. Концепція сховища даних. Організація сховища даних.
27. OLAP - системи. Багатовимірна модель даних.
28. Перетворення даних. Основні методи перетворення даних.
29. Квантування даних.
30. Злиття даних. Об’єднання, внутрішнє і зовнішнє з’єднання.

31. Угрупування даних. Приклади угрупування даних.
32. Поняття, задачі і роль візуалізації даних. Візуалізація 2D і 3D даних.
33. Візуалізація багатовимірних даних.
34. Очищення і попередня обробка даних.
35. Проста лінійна регресія.
36. Оцінення відповідності простої лінійної регресії реальним даним.
37. Модель множинної лінійної регресії.
38. Основи логістичної регресії.
39. Наївний Байєсовський класифікатор. Приклад роботи.
40. Поняття дерева рішень. Принципи побудові дерев рішень.
41. Оцінювання якості розбиття у деревах рішень. Основні проблеми побудови дерев рішень.
42. Критерії вибору кращих алгоритмів розбиття.
43. Алгоритми ID3 побудови дерев рішень. Приклад.
44. Поняття кластеризіції. Основні цілі кластеризації. Міри близькості об’єктів.
45. Ієрархічна кластеризація.
46. Алгоритм кластеризції k-means. Приклад.
47. Поняття мережі Кохонена. Алгоритм навчання мережі Кохонена.
48. Поняття карти Кохонена. Методика побудови. Види. Переваги і недоліки.
49. Ансамблі моделей.
50. Оцінення ефективності та порівняння моделей.
51. Поняття асоціативного правила. Основні характеристики асоціативних правил.
52. Алгоритм Apriori пошуку асоціативних правил.
53. Поняття послідовних шаблонів. Задача пошуку послідовних шаблонів.
54. Загальний алгоритм пошуку послідовних шаблонів.
55. Загальна постановка задачі пошуку. Два основних типи пошуку.
56. Принципи побудови еволюційних алгоритмів.
57. Поняття генетичного алгоритму. Структура генетичного алгоритму.
58. Класичні операції генетичного алгоритму. Приклади модифікації класичних операцій.
59. Проблема кодування інформації для генетичного алгоритму.
60. Функція пристосованості та використання механізму «рулетки» в генетичних алгоритмах.

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

Базова

1. Паклин Н. Б. Бизнес-аналитика: от данных к знаниям (+CD) : учебное пособие / Н. Б. Паклин, В. И. Орешков. – [2-е изд., испр.]. – СПб. : Питер. – 2013. – 704 с.
2. Анализ данных и процессов : учеб. пособие / [Барсегян А. А., Куприянов М. С., Холод И. И. и др.]. – [3-е изд., перераб. и доп.]. – СПб : БХВ-Петербург, 2009. – 512 с.
3. Черняк О. І. Інтелектуальний аналіз даних : підручник / О. І. Черняк, П. В. Захарченко. – К. : Знання, 2014. – 599 с.
4. Волкова П. А. Статистическая обработка данных в учебно-исследовательских работах : учебное пособие / П. А. Волкова, А. Б. Шипунов. – М. : ФОРУМ, 2012. – 96 с.
5. Грас Джоэл. Data Science. Наука о данных с нуля / Грас Джоэл. – СПб. : БХВ-Петербург, 2017. – 324 с.
6. Шатт Рэйчел Data Science. Инсайдерская информация для новичков / Шатт Рэйчел, О'Нил Кэти. – СПб. : Питер, 2019. – 368 с.
7. Силен Дэви Основы Data Scince і Big Data. Pyton и наука о данных / Силен Дэви, Мейсман Арно, Али Мохаммед. – СПб. : Питер, 2017. – 336 с.

Електронні ресурси

1. Інтелектуальний аналіз даних [Електронний ресурс] : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – Режим доступу:
   http://eir.zntu.edu.ua/ bitstream/123456789/ 2155/1/ – Назва з екрана.
2. Інтелектуальний аналіз даних. Комп'ютерний практикум [Електронний ресурс] : навч. посіб. для студ. спеціальності 122 «Комп'ютерні науки та інформаційні технології» / О. О. Сергеєв-Горчинський, Г. В. Іщенко 2012. – Режим доступу:
   https://ela.kpi.ua/bitstream/123456789/24971/1/ Komp_prakt.pdf
3. Технология Data Mining: Интеллектуальный анализ данных [Электронный ресурс] : конспект лекций. Режим доступу: http:// kek.ksu.ru/EOS/dm.pdf (дата 13.11.2019). – Назва з екрана.

ЛІТЕРАТУРА

1. Сбор данных – Data mining [Електронний ресурс]. – Режим доступу: https://ru.qaz.wiki/wiki/Data_mining
2. 100 Task [Електронний ресурс]. – Режим доступу: https://100task.ru/sample/27.aspx
3. Паклин Н. Б. Бизнес-аналитика: от данных к знаниям (+CD) : учебное пособие / Н. Б. Паклин, В. И. Орешков. – [2-е изд., испр.]. – СПб. : Питер. – 2013. – 704 с.

Навчальне видання

Методичні вказівки до виконання контрольних робіт з дисципліни «Інтелектуальний аналіз даних» для студентів заочної форми навчання спеціальності 122 – «Комп'ютерні науки»

Укладачі: Володимир Іванович Месюра
Ярослав Володимирович Іванчук
Олег Костянтинович Колесницький

Рукопис оформив *В. Месюра*

Редактор *Т. Старічек*

Оригінал-макет виготовила *Т. Криклива*

Підписано до друку 02.03.2021 р.
Формат 29,7×42 ¼. Папір офсетний.
Гарнітура Times New Roman.
Друк різографічний. Ум. друк. арк. 2,52.
Наклад 40 (1-й запуск 1-21) пр. Зам. № 2021-018.

Видавець та виготовлювач
Вінницький національний технічний університет,
інформаційний редакційно-видавничий центр.
ВНТУ, ГНК, к. 114.
Хмельницьке шосе, 95,
м. Вінниця, 21021.
Тел. (0432) 65-18-06.
press.vntu.edu.ua;
E-mail: kivc.vntu@gmail.com
Свідоцтво суб'єкта видавничої справи
серія ДК № 3516 від 01.07.2009 р.