

**Київський національний університет імені Тараса Шевченка**

**Економічний факультет**

**Кафедра економічної кібернетики**

**КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА  
«РОЗВИТОК РЕКОМЕНДАЦІЙНИХ СИСТЕМ З ВИКОРИСТАННЯМ  
МАШИННОГО НАВЧАННЯ»**

студентки 4 курсу

спеціальності 051 «Економіка»

ОПП «Економічна кібернетика»

денної форми навчання

Лем Тетяни Олександрівни

Науковий керівник:

доктор економічних наук, професор

Чорноус Галина Олександрівна

Засвідчую, що в цій роботі немає запозичень із  
праць інших авторів без відповідних посилань

Студент \_\_\_\_\_

Роботу допущено до захисту перед ЕК  
рішенням кафедри економічної кібернетики  
від 9 червня 2022 року, протокол № 15

Завідувач кафедри:

доктор економічних наук, професор

Ляшенко Олена Ігорівна \_\_\_\_\_

КИЇВ – 2022

## ЗМІСТ

ЗМІСТ .....	2
РЕФЕРАТ .....	3
RESUME.....	4
ВСТУП.....	5
РОЗДІЛ 1. ТЕОРЕТИЧНІ АСПЕКТИ ФУНКЦІОНУВАННЯ ТА РОЗРОБКИ РЕКОМЕНДАЦІЙНИХ СИСТЕМ.....	8
1.1. Особливості сучасних рекомендаційних систем .....	8
1.2. Огляд сучасного стану розвитку гібридних рекомендаційних систем .....	12
РОЗДІЛ 2. МАТЕМАТИЧНЕ ТА ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ РЕКОМЕНДАЦІЙНИХ СИСТЕМ.....	18
2.1. Огляд методів оцінювання продукції та розроблення моделі альтернативної оцінки.....	18
2.2. Огляд релевантних алгоритмів машинного навчання.....	25
2.3. Характеристики бази даних для розробки моделей рекомендацій.....	30
РОЗДІЛ 3. ПРАКТИЧНІ АСПЕКТИ РОЗРОБЛЕННЯ МОДЕЛЕЙ ПІДТРИМКИ РЕКОМЕНДАЦІЙ .....	33
3.1. Концептуальні основи розробки моделей підтримки рекомендацій .....	33
3.2. Обґрунтування необхідності використання скоригованих оцінок .....	34
3.3. Реалізація моделювання підтримки рекомендацій .....	37
ВИСНОВКИ.....	48
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ .....	50

## РЕФЕРАТ

Кваліфікаційна робота бакалавра містить: 54 сторінки, 24 рисунки, 11 таблиць, 45 джерел.

Ключові слова: рекомендаційна система, машинне навчання, електронна комерція, методи оцінювання в рекомендаційних системах.

Об'єкт дослідження: рекомендаційні системи в електронній комерції.

Мета дослідження: розробка альтернативного підходу до оцінювання продукції в рекомендаційних системах та пропозиція моделей підтримки рекомендацій з використанням алгоритмів машинного навчання.

Методи дослідження: загальнонаукові методи дослідження, методи аналізу, синтезу, порівняльного дослідження, економіко-математичного моделювання: методи побудови й аналізу моделей оцінювання продукції в рекомендаційних системах та їх програмна реалізація в середовищі R Studio, алгоритми машинного навчання в рекомендаційних систем та їх реалізація засобами аналітичної платформи IBM SPSS Modeler.

Наукова новизна, теоретична значимість дослідження: розроблено та реалізовано моделі підтримки рекомендацій, що використовують новий удосконалений метод оцінювання продукції з поєднанням машинного навчання.

Практична цінність: розширення кола аналізованих даних продукції та користувача для формування досконаліших персональних рекомендацій; розробленні формування оцінок продукції на базі більшої кількості метрик для покращення точності; вдосконаленні поєднання алгоритмів машинного навчання для моделей підтримки рекомендацій; у можливості використання розробленої концепції моделей на ринках електронної комерції з невисоким рівнем реалізації рекомендаційних технологій, зокрема в Україні.

## **RESUME**

Taras Shevchenko National University of Kyiv,

Faculty of Economics, Department of Economic Cybernetics

Key words: recommendation system, machine learning, e-commerce, product evaluation methods in recommendation systems

The graduation research of student Lem Tetiana deals with developing an alternative approach to product evaluation in recommendation systems and proposing a model of recommendation system using machine learning algorithms.

The work is interesting for the possibility of using the developed concept of a hybrid system with machine learning in markets with a low level of implementation of recommended technologies due to expanding the range of metrics and analyzed user data to provide better personal recommendations and high accuracy.

Pages 54, tables 11, figures 24, bibliographies 45.

## ВСТУП

**Актуальність роботи.** Розвиток електронної комерції, зростання кількості відповідних суб'єктів господарювання в глобалізованому економічному просторі загострює проблему їх конкурентоспроможності. Її вирішують за рахунок оптимізації рекомендаційних систем і їх гібридизації, що пов'язана з намаганням подолати певний недолік. При цьому менше уваги приділяється методології оцінювання продукції, тому існує нагальна необхідність в формуванні альтернативних систем оцінювання, що враховували б особливості рекомендаційних систем, виходячи з рівня реалізації рекомендаційних технологій. Для ринків з невисоким рівнем розвитку таких технологій (зокрема українського) характерне використання обмеженої інформації відстеження та розрахунок оцінок продукції за найпростішими метриками. Виходячи з цього для підвищення результативності рекомендаційних систем необхідне використання моделей підтримки рекомендацій відмінних від найбільш поширених та простих. Зазначене обумовлює актуальність досліджень, результатом яких має стати розробка моделей підтримки рекомендацій з удосконаленою методикою оцінювання продукції, що дозволяє підвищити конкурентоспроможність бізнесу в умовах невисокого рівня розвитку рекомендаційних технологій (відстежуваних даних) з використанням алгоритмів машинного навчання.

Проблемам розвитку рекомендаційних систем присвячено роботи багатьох іноземних вчених, серед яких Фен Дж., Фен Х., Пен Дж., Раджешвара Р., Тіаго С., Чжан М., Абід У., Ашраф М., Батт М., Чигозірім А., Патіл М., Рао М., Чжоу Л., Аніта Дж., Ріші О., Чжао Ц., Чжан І. та інші. Важливим є внесок у розвиток методів та засобів прогнозування рекомендацій також і вітчизняних дослідників, таких як Чорноус Г., Вишинський М., Харламова Г., Столарчик П., Плєскач В., Негрей М., Гнот Т., Шварц М., Замятін Д.

**Мета і задачі дослідження.** Метою наукової роботи є розробка альтернативного підходу до оцінювання продукції в рекомендаційних системах

та пропозиція моделей підтримки рекомендацій з використанням алгоритмів машинного навчання.

Відповідно до мети сформовано і вирішено ряд наступних завдань:

- окреслити ключові особливості сучасних гібридних систем;
- проаналізувати існуючі підходи до формування оцінки продукції;
- дослідити застосовані алгоритми машинного навчання в рекомендаційних системах;
- розробити ефективну альтернативну модель оцінювання;
- проаналізувати ефективність розробленої моделі оцінювання;
- розробити моделі підтримки рекомендацій з застосуванням машинного навчання.

Об'єктом дослідження є рекомендаційні системи в електронній комерції.

Предметом дослідження є методи оцінювання продукції та алгоритми машинного навчання в рекомендаційних системах.

**Методи дослідження.** Для вирішення поставлених завдань у роботі використано загальнонаукові методи дослідження, методи аналізу, синтезу, порівняльного дослідження, економіко-математичного моделювання: методи побудови й аналізу моделей оцінювання продукції в рекомендаційних системах та їх програмна реалізація в середовищі R Studio, алгоритми машинного навчання в рекомендаційних систем та їх програмна реалізація в середовищі IBM SPSS Modeler.

Інформаційною базою при написанні роботи виступали дисертаційні роботи, монографії, наукові статті вітчизняних та зарубіжних вчених, оприлюднені у фахових періодичних виданнях, набір даних Book-Crossing Dataset. Для реалізації моделей використано програмне забезпечення R Studio, IBM SPSS Modeler та Microsoft Access.

**Наукова новизна одержаних результатів** полягає у тому, що розроблено та реалізовано моделі підтримки рекомендацій, що використовують новий удосконалений метод оцінювання продукції з поєднанням алгоритмів машинного навчання.

**Практичне значення одержаних результатів** полягає у розширенні кола аналізованих даних для формування досконаліших персональних рекомендацій; розробленні формування оцінок продукції на базі більшої кількості метрик для покращення точності; вдосконаленні поєднання алгоритмів машинного навчання для моделей підтримки рекомендацій; у можливості використання запропонованого підходу на ринках електронної комерції з невисоким рівнем реалізації рекомендаційних технологій, зокрема в Україні.

**Апробація результатів дослідження.** Основні положення кваліфікаційної роботи апробовані на міжнародній науково-практичній конференції «Information Technology and Interactions (IT&I-2021)» (Київ, 2021 р.). За результатами доповіді опубліковано статтю за проблематикою дослідження: «Product valuation modeling in hybrid recommendation systems» (2021 р.), що індексовано у Scopus. Положення роботи також висвітлені у статті «Developing hybrid recommendation systems: Ukrainian dimension» у науковому періодичному іноземному фаховому виданні «Access to science, business, innovation in the digital economy» (Болгарія) та науковій роботі «Моделювання оцінки продуктів в гібридних рекомендаційних систем» (Київ, 2022 р.), що стала переможцем 1 туру конкурсу студентських наукових робіт зі спеціальності «Економічна кібернетика» у 2021/2022 н.р.

**Структура та обсяг роботи.** Наукова робота складається зі вступу, трьох розділів, висновків, списку використаних джерел і додатків. Перший розділ роботи зосереджений на виділенні ключових рис рекомендаційних систем та виділенні стану розвитку рекомендаційних систем в розрізі існуючий робіт. У другому розділі розглянуто математичний апарат та базу даних, що застосовувалися для реалізації моделювання. Третій розділ зосереджений на описі, особливостях та ефективності застосування методики оцінювання продукції та алгоритмів машинного навчання в моделях підтримки рекомендацій. Загальний обсяг роботи становить 54 сторінок тексту (49 сторінок основної частини), включаючи 11 таблиць, 24 рисунки та 45 найменування у списку використаних джерел.

## **РОЗДІЛ 1. ТЕОРЕТИЧНІ АСПЕКТИ ФУНКЦІОНУВАННЯ ТА РОЗРОБКИ РЕКОМЕНДАЦІЙНИХ СИСТЕМ**

### **1.1. Особливості сучасних рекомендаційних систем**

Рекомендаційні системи - це експертні системи фільтрації інформації, що мають на меті передбачити потенційні потреби користувачів шляхом аналізу їх уподобань та надання персоналізованих рекомендаційних послуг. Їх використання допомагає споживачам відкрити нові види продукції, що можуть бути їм цікаві [1, 2]. В інтересах як платформ (бізнесу), так і споживачів застосовувати такі рекомендаційні системи, що збільшують новизну та різноманітність списку рекомендацій. Продукти, щодо яких найчастіше використовуються рекомендаційні системи є фільми, книги, новини, дослідницькі статті, товари та предмети розкоші.

Історія розвитку рекомендаційних систем розпочалася в галузі науки пізнання та пошуку інформації, її першим проявом стала комунікаційна система Usenet, створена Університетом Дьюка в другій половині 1970-х років, де користувачі могли обмінюватися текстовим вмістом з розподілом на групи новин і підгрупи для спрощення пошуку. З часом розвивалися два дуже різних напрямки рекомендаційних систем: спільна фільтрація та фільтрація на основі вмісту. Першим прикладом спільної фільтрації була система Tapestry, розроблена Херох PARC, яка дозволяла своїм користувачам робити нотатки та коментувати документи у двійковій формі. Одним із перших і досить успішних розробок рекомендаційної системи на основі вмісту був проект «Музичний геном» у 1999 році, який мав на меті «зрозуміти» та охопити музику через її властивості. Першим рішенням щодо гібридної рекомендаційної системи був Fab у 1994 році, що частково усував недоліки окремих двох методів. Ця модель збирала контент з певних тем, а потім для кожного окремого користувача вибирала ті елементи, які ймовірніше за все, їх зацікавлять [3, 4].



Перший етап рекомендаційної системи (рисунок 1.1) полягає в зборі та накопиченні інформації про користувачів і елементи для формування відповідних профілів. На другому етапі використовуються методи навчання та алгоритми для обробки інформації користувача, отриманої за допомогою зворотного зв'язку. Останній пункт — використання зворотного зв'язку. Явний зворотний зв'язок полягає в тому, що користувач коректує прогнозовані рейтинги предметів після їх використання. При неявному зворотному зв'язку система автоматично визначає уподобання користувача шляхом відслідковування різних дій користувача [5].



Рис. 1.1. Узагальнена архітектура рекомендаційної системи

Джерело: [5]

Можна виділити 7 методів для реалізації рекомендаційних систем: на основі вмісту (контенту); залежний від випадку; залежний від атрибутів; спільна фільтрація; демографічна рекомендація; рекомендація, залежна від знань; гібридна система рекомендацій; рекомендація залежна від онтології [6].

Основними в використанні в рекомендаційних системах є методи на основі вмісту, спільної фільтрації та гібридні методи.

Системи рекомендації на основі вмісту включають роботу з профілем конкретного користувача. Для рекомендацій певного елемента враховуються множина описів властивостей попередніх елементів, а також їх оцінки, вподобання, теги, ключові слова. Результатом є оцінка рівня релевантності, яка визначає рівень інтересу користувача до предмета [5]. Проблема використання таких систем присвячені праці [7, 8, 9].

Системи рекомендацій на основі спільної фільтрації базуються на пошуку та аналізі минулої поведінки користувача, на основі якої згодом прогнозуються елементи на подібності типу рейтингів, що даються користувачами-однодумцями цільовому користувачу. Більшість існуючих методів спільної фільтрації значною мірою покладаються на явні дані зворотного зв'язку. Основним припущенням методу є те, що користувачі, які мали певні уподобання в минулому, будуть мати такі ж в майбутньому. Методи спільної фільтрації можна класифікувати на такі, що ґрунтуються на пам'яті та моделі [10]. Результати досліджень щодо таких систем представлено в [11, 12, 13, 14].

Гібридні рекомендаційні системи поєднують методи спільної фільтрації і методи на основі вмісту. На основі досліджень [1, 5, 15, 16] можна виділити сім основних принципів, які демонструють різні способи поєднання методів на основі змісту та спільної фільтрації у гібридну систему рекомендацій:

- зважування – поєднання компонентів різних типів рекомендаційних систем з певними вагами для створення спільного одного результату;
- перемикання – зміна типу системи залежно від наявної ситуації;
- мікс - поєднання результатів різних типів рекомендаційних систем;
- комбінація особливостей - функції з різних типів систем поєднуються і подаються як вхідні дані до єдиного алгоритму.
- нарощування особливостей - використання методів одного типу системи для обчислення ознак, що є частиною вхідних даних іншої системи;

- каскадний - результат системи одного типу виступає як вхідний сигнал до методики рекомендаційної системи іншого типу;
- мета рівень - техніка одного типу системи створює певну модель, яка потім є вкладом техніки рекомендаційної системи іншого типу.

В межах створення гібридної рекомендаційної системи можливе різне поетапне поєднання методів різного типу рекомендаційних систем. Найбільш поширеними є паралельна та монолітна побудова [17]. Монолітний тип має наступні ознаки: лише один компонент рекомендації; гібридизація є "віртуальною" в тому сенсі, що поєднуються ознаки/джерела знань рекомендаційних систем різних типів. Паралельний тип має наступні ознаки: результат сформований на основі кількох існуючих реалізацій різних типів рекомендаційних систем; найменш інвазійний дизайн; присутня схема зважування або голосування для поєднання кількох джерел знань.

Система гібридних рекомендацій інтегрує різні типи, що допомагає подолати недоліки та покращити продуктивність функціонування. Емпіричні дослідження показують вищу результативність та якість гібридного підходу і доводять, що такі підходи надають рекомендації точніше, ніж підходи на основі вмісту чи спільної фільтрації [15, 7, 18, 19, 20]. Більшість сучасних систем створені для усунення таких найпоширеніших проблем, як:

- а) Обмежений аналіз вмісту - використовується явний опис характеристик предметів, що не змінюється в процесі роботи.
- б) Надмірна спеціалізація - обмеження користувачів елементами, подібними до обраними ними, а отже, нові елементи і варіанти не виявляються.
- в) Холодний старт - новий предмет або новий користувач. Рекомендація є складнішою для аналізу, оскільки користувачі ще не оцінили нові предмети/недостатньо інформації про уподобання нового користувача.
- г) Масштабованість – стик з проблемами в обробці величезної кількості інформації, в результаті надана рекомендація може бути неточною.
- д) Розрідженість - матриця оцінок є розрідженою у випадку, коли більшість користувачів не беруть участі в оцінці предметів [17].

Ефективність рекомендаційних алгоритмів вимірюють, використовуючи механізми їх оцінювання. Складність оцінки рекомендованої системи залежить від кількості підходів або алгоритмів. Найбільш поширені метрики оцінок в рекомендаційних системах: рейтинг і використовувана точність прогнозу, охоплення користувача та простору, температура користувача та товару, ранжування, розширені метрики та онлайн метрики.

## 1.2. Огляд сучасного стану розвитку гібридних рекомендаційних систем

У цьому параграфі розглянуто прогресивні сучасні приклади гібридних рекомендаційних систем. Дослідження представлених систем дозволяє проаналізувати стан і вектор розвитку рекомендаційних систем загалом. З іншого боку, дослідження цих систем допоможе при формуванні моделей підтримки рекомендацій. Представлено результати огляду досліджень рекомендаційних систем з поділом на дві групи. Роботи в першій з них зосереджені на методах оцінювання продукції в гібридних рекомендаційних системах. В другій групі досліджень виділено застосування машинного навчання в рекомендаційних системах.

### Огляд рекомендаційних систем з акцентом на метод оцінювання

*Гібридна система з подоланням проблеми холодного старту користувача.*

У роботі [1] автори розглянули крайній випадок розріджених даних, а саме проблему холодного запуску нового користувача. Для подолання цієї проблеми запропоновано модель рейтингу CF, яка поєднує рейтинговий підхід імовірнісної матриці факторизації (PMF) та попарно орієнтований на рейтинг підхід Байєсового персоналізованого рейтингу (BPR) разом. Представлена модель повністю використовує явні та неявні дані зворотного зв'язку.

*Гібридна система з зосередженням на довгому хвості.* Дослідження [10] враховує вплив вподобань/смаків окремих споживачів щодо нових та різноманітних предметів. У роботі використано розширення моделі факторизації матриць для прогнозування рейтингу. Запропоновано моделі (PM-1, PM-2), що демонструють придатність: 1) надавати персоналізовані рекомендації

користувачеві, беручи до уваги відповідний смак предметів з довгим хвостом; 2) рекламувати предмети з довгим хвостом до ідіосинкратичних користувачів .

*Гібридна система на основі рекомендацій та тональності.* У роботі [15] запропоновано рекомендаційна система, яка базується на гібридному аналізі рекомендацій та тональності з використанням F-метрики. Запропонована у дослідженні модель є ефективнішою у своїй здатності ідентифікувати відповідні фільми користувачу. Показники, які аналізує система: відгуки, опубліковані користувачами; характеристики рекомендацій фільмів, історія перегляду користувачів; позитивний та негативний загальний тренд відгуків.

*Гібридна система для рекомендацій активним користувачам.* Автори дослідження [17] побудували рекомендаційну систему NLM (novel recommendation system), що поєднує методи спільної фільтрації та матриці факторизації. NLM використовує генетичні алгоритми для оцінки ставок неоцінених предметів активного користувача. Розглянута рекомендаційна система для спільної фільтрації використовує моделі сусідства і моделі прихованого фактору, щоб рекомендувати елементи. Метод факторизації використовується в моделі прихованих факторів для пошуку високо очікуваних рейтингових позицій, активних користувачів, які мають високу перевагу.

*Гібридна система для пошуку навчальних матеріалів.* У дослідженні [18] пропонується вдосконалений метод для існуючої системи рекомендацій електронного навчання із поєднанням фільтрації на основі вмісту та спільної фільтрації з хорошими рейтингами учнів (метод CBF-CF-GL). Спосіб прийняття для аналізу учнів лише з високими балами може бути використаним в електронній комерції, наприклад, через відбір для аналізу користувачів з витратами в певному конкретному інтервалі, що релевантний до витрат цільового користувача.

*Гібридна система з урахуванням коефіцієнтів подібності профілів та демографічних відмінностей.* У праці [5] розроблено метод розрахунку коефіцієнта подібності власних векторів користувача та власних векторів об'єкта, який використовує демографічні характеристики користувача.

Застосовано алгоритм пошуку асоціативних правил Апріорі за допомогою адаптивної зміни підтримки асоціативних правил. Наведені результати експериментального дослідження коефіцієнтів подібності, які обчислені за конусною мірою подібності, коефіцієнтом кореляції Пірсона, коефіцієнтом Жаккарда і оберненою евклідовою відстанню, коефіцієнтом подібності, який враховує демографічні характеристики користувачів.

*Гібридна система з методом часткового варіаційного автокодеру.* Авторами [10] запропоновано метод гібридної рекомендації - частковий варіаційний автокодер, який обробляє відсутні дані та використовує амортизований висновок для швидкого прогнозування. Метод використовує нову імовірнісну генеративну модель для обробки різної кількості рейтингів користувачів. Використовуючи запропоновану амортизовану техніку часткового висновку в р-VAE, навчання та умовивід можуть бути ефективно виконані шляхом мінімізації часткової варіаційної верхньої межі, не роблячи спеціальних припущень щодо значень відсутніх оцінок.

#### Огляд рекомендаційних систем з застосуванням машинного навчання

*Система рекомендацій, що зосереджена на оптимізації навігації та підсумовуванні огляду продукту за допомогою машинного навчання та методів штучного інтелекту.* Робота авторів [21] розробила систему рекомендацій на основі демографічного вмісту спільно з використанням міри гібридної схожості. Оптимізація навігації здійснюється за допомогою алгоритму оптимізованого префіксного діапазону.. Застосовано структуру класифікатора латентного розподілу Діріхле на основі вибірки Гіббса, що використовується для класифікації відгуків про продукт на позитивні, негативні та нейтральні та представляє їх у вигляді стовпчастої діаграми. Використання такого підходу зменшуватиме людські зусилля під час здійснення покупок на сайті електронної комерції та сприятиме високоякісному досвіду користувачів із більшою відносною ефективністю та рівнем задоволення.

*Застосування глибокого навчання та технології розподіленого виразу в рекомендаційних системах електронної комерції.* У роботі [22] на семантичному

рівні реклами побудовано мережу подібності на основі розподілу реклами за темами, а на цій основі побудовано структуру моделі глибокого навчання для прогнозування рейтингу кліків реклами. Результатом є запропонований вдосконалений алгоритм рекомендацій, заснований на рекурентній нейронній мережі та розподіленому виразі. Це покращує традиційну рекурентну нейронну мережу та вводить часове вікно для керування передачею даних прихованого рівня рекурентної нейронної мережі. Така модель перевершує традиційну модель рекурентної нейронної мережі за точністю і зменшенням складності розрахунків.

*Рекомендаційна система з використанням машинного навчання на основі подібності* У дослідницькій роботі [23] застосовано техніку зменшення розмірності за допомогою аналізу головних компонентів (PCA) через розкладання сингулярних значень (SVD) для перетворення вилучених об'єктів у простір нижньої розмірності. Використано підхід до кластеризації K-Means++ для можливої ідентифікації кластера для подібної групи продуктів. Пізніше обчислено відстань від Манхеттена для вхідного зображення до цільових кластерів, встановлених для отримання перших N подібних продуктів із низькою мірою відстані. Підхід порівняно з п'ятьма різними неконтрольованими алгоритмами кластеризації, а саме Minibatch, K-Mediod, Agglomerative, Brich і Gaussian Mixture Model (GMM). Обчислено різні показники ефективності кластерів на K-середніх++ і досягнуто коефіцієнта силуету (SC) 0,1414, індексу Калінського-Харабаша (CH) 669,4 і індексу Девіса-Болдіна (DB) 1,8538. Запропонований PCA-SVD підхід, трансформований K-mean++, показав кращу продуктивність порівняно з іншими п'ятьма підходами до кластеризації для подібних зображень продуктів.

*Рекомендаційна система на основі механізму опорного вектору (SVM), що допомагає покращити проблеми в техніці спільної фільтрації.* Рекомендаційна система авторів [24] включає (1) SVM класифікатор для класифікації об'єктів на позитивний та негативний зворотний зв'язок. Найкраще досягнуте значення вказує на оптимізовані значення параметрів SVM з використанням алгоритму IACO, які надаються у вигляді вхідних даних до класифікатора для виконання

попарної класифікації. Згодом (2) використано алгоритм фільтрації на основі SVM–IACO. Тести даних Таобао показали, що алгоритм дає високу точність рекомендацій.

*Підхід на основі машинного навчання для подібної системи рекомендацій на основі зображень.* У дослідницькій роботі [25] застосовано техніку зменшення розмірності за допомогою аналізу головних компонентів (PCA) через розкладання сингулярних значень (SVD) для перетворення вилучених об'єктів у простір нижньої розмірності. В межах підходу використано кластеризацію методом K-Means++ для ідентифікації подібних груп продуктів. Також обчислено відстань від Манхеттена для вхідного зображення до цільових кластерів, встановлених для отримання перших N подібних продуктів із низькою мірою відстані. Запропонований PCA-SVD підхід, трансформований з K-mean++, показав кращу продуктивність порівняно з іншими п'ятьма підходами (Minibatch, K-Mediod, Agglomerative, Brich, Gaussian Mixture Model) до кластеризації для подібних рекомендацій щодо продуктів зображення.

*Рекомендаційна система, що використовує базу знань, згенеровану на основі графа знань, щоб ідентифікувати предметні знання користувачів, предметів та зв'язків між ними [26].* Граф знань є позначеним багатовимірним орієнтованим графом, який представляє відносини між користувачами та елементами. Запропонований підхід використовує близько 100% участі користувачів у формі діяльності під час навігації веб-сайтом. Запропонована система пов'язує категорію предметів, а не лише конкретні елементи, які можуть зацікавити користувачів. Ефективність цього підходу в порівнянні з базовими методами підтверджено за трьома параметрами (прецизійність, запам'ятовування і NDCG через онлайн та офлайн оціночні дослідження з даними користувачів).

*Причини та способи автоматичної реалізації встановлення знижки ціни у рекомендаційній системі [27].* Ключ до оптимізації знижки полягає у передбаченні готовності споживача платити (WTP), а саме визначенні найвищої ціни, яку споживач готовий заплатити за продукт. Запроваджений механізм



електронної комерції, адаптований на основі експериментів з лабораторних лотерей та аукціонів, які вимагають раціонального WTP клієнта для невеликої підмножини продуктів, і використаний алгоритм машинного навчання для прогнозування WTP клієнта для інших продуктів. Механізм реалізовано на нашому власному веб-сайті електронної комерції, який використовує дані та суб'єкти Amazon, залучені через Mechanical Turk. Підхід може допомогти передбачити WTP і підвищити задоволеність споживачів, а також прибуток продавця.

В роботі запропоновано моделі підтримки рекомендацій, що відрізняється від розглянутих покращеною точністю оцінок продукції за рахунок аналізу ширшого кола релевантних показників з відповідної нормалізацією. Розглянуті дослідження рекомендаційних систем з використанням машинного навчання дозволити сформувати базу використання алгоритмів для реалізації в рекомендаційній системі.

## РОЗДІЛ 2. МАТЕМАТИЧНЕ ТА ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ РЕКОМЕНДАЦІЙНИХ СИСТЕМ

2.1. Огляд методів оцінювання продукції та розроблення моделі альтернативної оцінки

Розроблення ефективних рекомендаційних систем потребує використання відповідного математичного апарату та моделювання. Середній рейтинг (проста рекомендація) є найбільш простим видом рекомендації. Використовується у тому випадку, коли обмежений час розрахунку. Отриманий рейтинг  $r_{av}$  є середнім арифметичним усіх значенням, обрахований за формулою (2.1) [28].

$$r_{av} = \frac{\sum_{j=1}^n r_{ij}}{n}, \quad (2.1)$$

де  $r_{av}$  - середній рейтинг продукту,

$r_{ij}$  – оцінка, виставлений кожним окремим користувачем,

$n$  – загальна кількість оцінок.

Для збалансування пропорції позитивних рейтингів з невизначеністю невеликої кількості спостережень одним з варіантів застосування математичного апарату є *розрахунок меж балу Вілсона довірчого інтервалу для параметра Бернуллі* за формулою (2.2) [28, 29].

$$r_w = \frac{\hat{p} + \frac{z^2 \frac{\alpha}{2}}{2n} \pm z \frac{\alpha}{2} \sqrt{\left[ \hat{p}(1 - \hat{p}) + \frac{z^2 \frac{\alpha}{2}}{4n} \right] / n}}{1 + \frac{z^2 \frac{\alpha}{2}}{2n}}, \quad (2.2)$$

де  $r_w$  – бал Вілсона,

$\hat{p}$  - спостережувана частка позитивних оцінок,

$z^2 \frac{\alpha}{2}$  квантиль стандартного нормального розподілу,

$n$  - загальна кількість оцінок.

Оцінка Вілсона буде тим вища, чим вища частка позитивних оцінок та чим більше оцінюваного продукту. Межі коливання від 0 (у випадку, коли немає виставлених оцінок або немає позитивних балів) до 1 (у випадку, коли велика кількість виставлених оцінок і всі вони позитивні).

За результатами аналізу продемонстровано (рисунок 2.1), що різке зростання оцінки відбувається при зростанні кількості оцінок в інтервалі від 0 до 100 (тим крутіший ріст, чим вища частина позитивних відгуків). При зростанні кількості виставлених оцінок понад 100 інтенсивність росту оцінки Вілсона є меншою. На рисунку продемонстровано загальну ситуацію, але межу кількість оцінок від якої скоригована оцінка не так різко зростає потрібно коригувати залежно від активності окремого сайту електронної комерції.

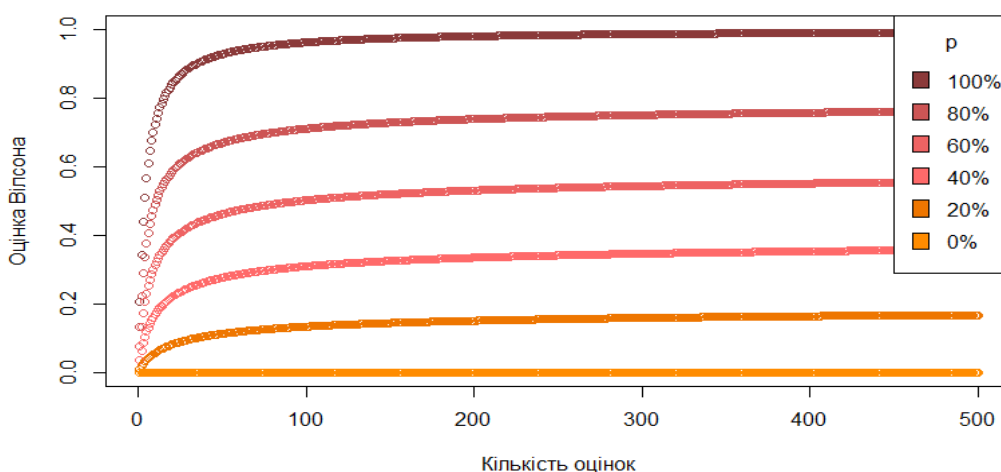


Рис. 2.1. Зміна оцінки Вілсона залежно від кількості оцінок (від 0 до 500) та частки позитивних оцінок

Джерело: сформовано автором

Іншою важливою тенденцією оцінки Вілсона є її однозначне зростання при збільшенні частки позитивних оцінок. Особливо яскраво це спостерігається при великій кількості оцінок. У випадку, коли кількість оцінок не є великою (наприклад до 20), зміна оцінки Вілсона є не такою різкою і більш плинною (рисунок 2.2). Так наприклад, оцінка Вілсона буде практично однаковою у

випадку, коли кількість оцінок рівна 5 і 80% з них позитивні до тієї, коли кількість оцінок рівна 2 і 100% позитивні.

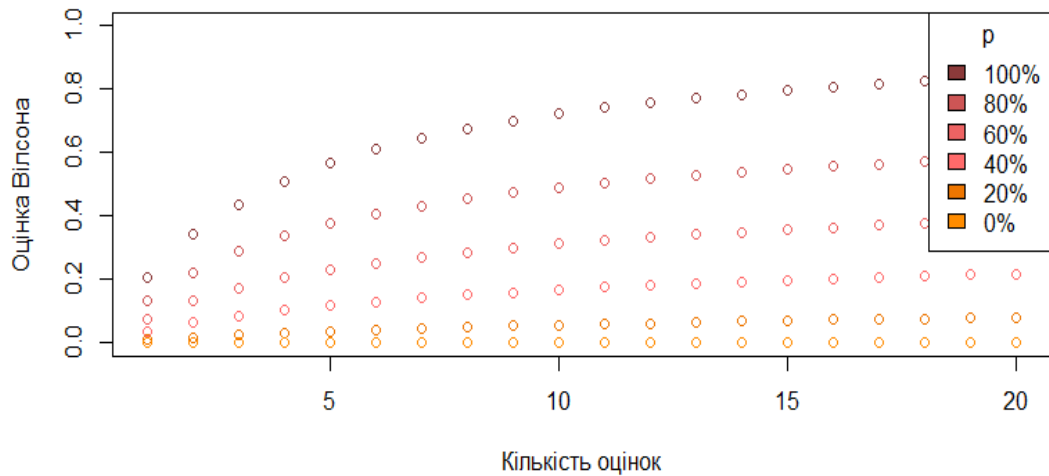


Рис. 2.2. Зміна оцінки Вілсона залежно від кількості оцінок (від 0 до 20) та частки позитивних оцінок

Джерело: сформовано автором

Іншою метрикою є *Байєсівська апроксимація*, що забезпечує спосіб оцінки товару, коли він вимірюється за шкалою  $K$  [29]. Метрика використовує нижню межу нормального наближення до Байєсівського довірчого інтервалу для середньої оцінки і розраховується за формулою (2.3).

$$r_b = \sum_{k=1}^K s_k \frac{n_k + 1}{N + K} - z_{\alpha/2} \sqrt{\frac{\left( \left( \sum_{k=1}^K s_k^2 \frac{n_k + 1}{N + K} \right) - \left( \sum_{k=1}^K s_k \frac{n_k + 1}{N + K} \right)^2 \right)}{N + K + 1}}, \quad (2.3)$$

де  $r_b$  - бал Байєса,

$s_k$  - 1-бальна, 2-бальна, ...,  $N$ -бальна шкала,

$N$  - загальна оцінка з оцінками  $n_k$  для шкали  $k$ .

На відміну від оцінки Вілсона, оцінка Байєса не проводить поділ на позитивні та негативні, а за основу бере кількість балів кожного виду. Чим більша кількість оцінок і чим більше з них найвищого балу, тим швидше оцінка наближається до останнього, водночас оцінка Байєса завжди більше 0.

В таблиці 2.1 проаналізовано різні варіанти оцінки Байєса за 10-ти бальною шкалою. Згідно з даними можна зробити висновок, що зменшення загальної кількості оцінок не просто зменшує оцінку Байєса незалежно від внутрішнього розподілу, а пом'якшує її. Так наприклад, у випадку однозначного вибору оцінки 9 і 10 при загальній кількості оцінок 10000, 100, 10 оцінка Байєса відповідно дорівнює 9,48; 8,09; 2,65 – існує тенденція до зменшення. В протилежному випадку однозначного вибору оцінок 1 і 2 при загальній кількості оцінок 10000, 100, 10 оцінка Байєса відповідно дорівнює 1,50; 1,22; 0,04 – при невеликій кількості загальних оцінювань бал Байєса також є нижчим, ніж при великій кількості оцінок. В більш реальних ситуаціях – тяжіння до певного балу прямо пропорційно залежить від загальної кількості оцінок і тяжіння до певного балу.

Таблиця 2.1

Оцінки Байєса залежно від різної кількості оцінок кожного виду балу

Кількість оцінок 1 і 2	Кількість оцінок 3 і 4	Кількість оцінок 5 і 6	Кількість оцінок 7 і 8	Кількість оцінок 9 і 10	Загальна кількість оцінок	Оцінка Байєса
0	0	0	0	10000	10000	9,48
0	0	0	10000	0	10000	7,49
0	0	10000	0	0	10000	5,49
0	10000	0	0	0	10000	3,50
10000	0	0	0	0	10000	1,50
0	0	0	0	100	100	8,09
0	0	0	100	0	100	6,45
0	0	100	0	0	100	4,78
0	100	0	0	0	100	3,05
100	0	0	0	0	100	1,22
0	0	0	0	10	10	2,65
0	0	0	10	0	10	2,29

Продовження табл. 2.1

Кількість оцінок 1 і 2	Кількість оцінок 3 і 4	Кількість оцінок 5 і 6	Кількість оцінок 7 і 8	Кількість оцінок 9 і 10	Загальна кількість оцінок	Оцінка Байєса
0	0	10	0	0	10	1,80
0	10	0	0	0	10	1,09
10	0	0	0	0	10	0,04
1000	1000	1000	2000	3000	10000	5,26
1000	2000	2000	4000	1000	10000	5,81
1000	2000	4000	2000	1000	10000	5,41
1000	4000	2000	2000	1000	10000	5,01
3000	2000	1000	1000	1000	10000	3,28
10	10	10	20	30	100	3,79
10	20	20	40	10	100	4,66
10	20	40	20	10	100	4,35
10	40	20	20	10	100	3,96
30	20	10	10	10	100	2,22
1	1	1	2	3	10	0,90
1	2	2	4	1	10	1,55
1	2	4	2	1	10	1,47
1	4	2	2	1	10	1,29
3	2	1	1	1	10	0,36

Джерело: сформовано автором

Метрика *Хакер* першочергово розроблена для ранжування новин для виділення гарячих та нових статей за допомогою параметрів впливу сили тяжіння ( $g$ ) та часу ( $t$ ) згідно з формулою (2.4). Однак цей алгоритм може бути застосований і для електронної комерції в наступних напрямках врахуванні відрізка часу: 1) публікації продукції; 2) виробництва продукції; 3) оцінювання продукції користувачами [30].

$$r_h = \frac{a}{(t + 2)^g}, \quad (2.4)$$

де  $r_h$  - бал Хакера,

$a$  - бали за позицією;

$t$  - час з моменту подання;

$g$  – гравітація (за замовчуванням 1,8 в news.arc).

Найефективнішим способом використання оцінки Хакер в електронній комерції є врахування часу публікації товару при наявній оцінці. Однак метрика часу публікації не завжди відслідковується, особливо в умовах невисокого рівня розвитку рекомендаційних технологій. В аналізованій базі даних [31] відсутня змінна час публікації, натомість використовується змінна час випуску. Метод Хакер коригує конкретну оцінку, а не ряд в сукупності. При збільшенні ступеня важливості врахування часу публікації оцінка Хакер стає більш крутою - чутливою до тривалості (рисунок 2.3 – ступінь важливості 0,5). За 10-ти бальною шкалою при наявності часу виробництва оцінка Хакер є однаковою для товару з стандартною оцінкою 8 при тривалості публікації 2 дні та товару з стандартною оцінкою 10 при виробництві 4 роки тому. При зменшенні оцінки бал Хакер відповідно зменшується (існує пряма залежність).

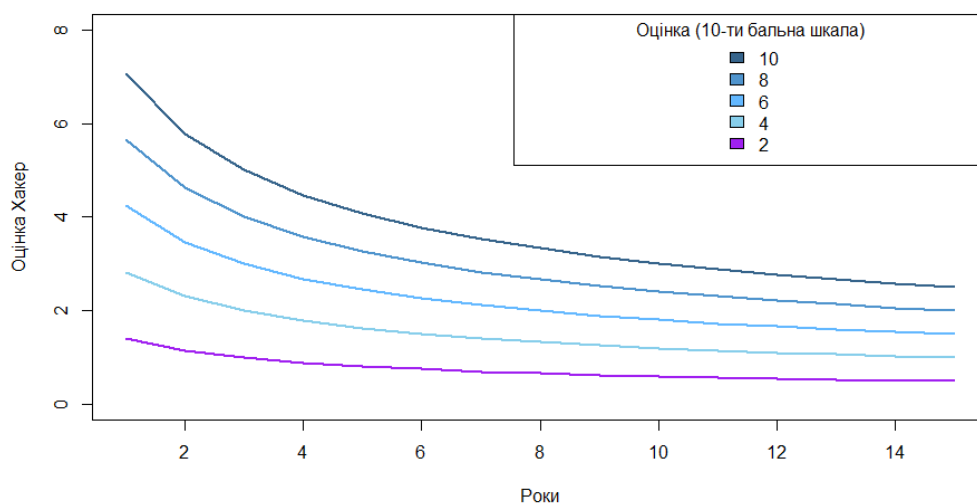


Рис. 2.3. Зміна оцінки Хакер залежно від кількості днів публікації

Джерело: сформовано автором

В роботі запропоновано оригінальну модель оцінки продукції на основі таких методів, як бал Вілсона, байєсівська апроксимація та алгоритму ранжування Хакер. Запропонована оцінка  $R$  розраховується за формулою (2.5).

$$R = \frac{a\sqrt{r_w^2 + b \times r_b^2}}{r_h}, \quad (2.5)$$

де  $R$  – оцінка за розробленою методикою;

$r_w$  – оцінка Вілсона;

$r_b$  – оцінка Байєса;

$r_h$  – оцінка Хакер,

$b$  – коефіцієнт нормалізації;

$a$  – коефіцієнт розмаху оцінки.

В розрахунку оцінки  $R$  поєднано метрики Вілсона та Байєса як середньоквадратична величина значень їх нормуванням. В роботі коефіцієнт  $b$  нормує значення цих оцінок між собою для забезпечення однакової верхньої межі оцінок Байєса і Хакера. У випадку, коли розробник надає перевагу певному методу оцінки, коефіцієнт  $b$  коригується. З методу Хакер взято коригування на обмеження часу виробництва продукції; а замість середньої оцінки продукції взято середньоквадратичне значення з нормування метрик Вілсона та Байєса. Коефіцієнт  $a$  встановлюється користувачем лише для розширення розмаху оцінки і не впливає на суть розрахунку. В роботі межі розробленої оцінки від 0 до 10. Розроблена модель має застосування для різних категорій продукції електронної комерції. Тому припускається, що давніший рік виробництва продукції зменшує привабливість для клієнтів. Оскільки скоригована оцінка розроблена для електронної комерції, ступінь коригування часу повинна бути меншою, ніж для новин. Така оцінки бере переваги від кожного з методів та коригує значення оцінки, якщо за одним з методів оцінка була б рівна 0. Це дозволяє сформувати оцінку середню по усій вибірці, і враховуючи поділ на позитивні та негативні оцінки (припускається, що оцінки 10, 9, 8, 7, 6, 5 – позитивні, оцінки 4, 3, 2, 1, 0 – негативні).



З метою визначення точності прогнозу в алгоритмах рекомендаційних систем обраховуються помилки (відхилення між прогнозованим і реальним рейтингом) [32]. Найбільш використовуваними метриками є середня абсолютна помилка (MAE), середньоквадратична помилка (MSE) і корінь середньоквадратичної помилки (RMSE), що розраховані за формулами (2.6-2.8) відповідно [33, 34].

$$MAE = \frac{1}{|Q|} \sum_{(u,i) \in Q} |r_{ui} - \hat{r}_{ui}|, \quad (2.6)$$

$$MSE = \frac{1}{|Q|} \sum_{(u,i) \in Q} (r_{ui} - \hat{r}_{ui})^2, \quad (2.7)$$

$$RMSE = \sqrt{\frac{1}{|Q|} \sum_{(u,i) \in Q} (r_{ui} - \hat{r}_{ui})^2}, \quad (2.8)$$

де  $MAE$  - середня абсолютна помилка,

$MSE$  - середньоквадратична помилка,

$RMSE$  - корінь середньоквадратичної помилки,

$Q$  - тестовий набір,

$r_{ui}$  - справжні оцінки користувача,

$\hat{r}_{ui}$  — прогнозовані рейтинги системи рекомендацій.

## 2.2. Огляд релевантних алгоритмів машинного навчання

В даній частині роботи розглядаються окремі алгоритми машинного навчання, застосування яких покращує якість рекомендацій, що надають системи. Алгоритми машинного навчання можна представити з поділом на наступні класи залежно від поставлених задач:

- Класифікація. Застосовується для розв'язання проблем, пов'язаних з присвоєнням класів кожному з аналізованих елементів, використовуючи навчання з вчителем. Прогнозуються предиктори як вхідні дані для отримання значення залежної змінної. Процес класифікації складається з етапів, таких як конструювання моделі та її подальше використання. Варіантом застосування в

рекомендаційних системах є розподіл продукції на групи, що подібні за метриками, що збирає сайт електронної комерції.

- Регресія. Метод, що визначає силу і характер зв'язку між однією залежною змінною і набором незалежних змінних. В рекомендаційних системах застосування регресійної залежності можливе для прогнозування оцінки продукції і визначені нетривіальних показників, що впливають на неї.

- Рейтинг. Сортуювання показників / продуктів за певним критерієм. Застосування можливе для сортування продукції за визначеними на основі аналізу більш вагомими метриками в рекомендаційних системах.

- Кластеризація. Розглядається поділ об'єктів на певні однорідні групи (кластери), використовуючи навчання без вчителя. Кластерний аналіз не накладає обмежень на властивості об'єктів. Його застосування в рекомендаційних системах є можливим для різних категорій об'єктів, великого обсягу даних при його подальшій оптимізації (стисненні, скорочуванні, підвищенні наочності), виявлення аномалій. Прикладом є виділення кластерів користувачів, що на основі активності на сайті, є подібними.

- Зменшення розмірності Перетворення вихідного представлення досліджуваного об'єкта в представлення менших розмірів без втрати вихідних даних, використовуючи кореляційні залежності. Для рекомендаційних систем це дозволяє залишати для аналізу лише найбільш значущі показники. Наприклад, для окремих алгоритмів достатньо і ефективніше брати до уваги кількість позитивних та негативних оцінок без загальної кількості оцінок.

Алгоритми машинного навчання водночас підпадають під поділ залежно від типу:

- Навчання під керівництвом (навчання з вчителем). Алгоритм отримує тренувальні дані, у яких виокремлене вихідне значення, відоме з вхідних даних.

- Навчання без нагляду (навчання без вчителя). На відміну від методу навчання з викладачем, алгоритм отримує тренувальні дані, які не враховують, а вихідне значення отримується з вхідних даних.

- Навчання з напівнаглядом. Тренувальні дані, хоча і частково контрольовані, складаються із вибірок, які мають очікуване початкове значення, а також вибірок, які його не мають.

- Навчання з підкріпленням. Фази навчання та тестування поєднуються у підході з підкріпленням. Вивчений алгоритм, взаємодіючи з середовищем, збирає дані. Він отримує, залежно від вчиненої дії, винагороду або штраф. Метою є максимізація винагороди за вивчений алгоритм [35, 36].

Нижче розглянуто окремі алгоритми машинного навчання, що можуть бути використані при формуванні моделей підтримки рекомендацій в рекомендаційних системах.

Алгоритм *K-середніх* є прикладом кластерного аналізу і має на меті розбити набір спостережень на визначену кількість кластерів (рисунки 2.4). Основні положення методу описуються наступним алгоритмом [37]:

- К точок розміщуються в просторі даних об'єкта, що представляє початкову групу центроїдів.
- Кожен об'єкт або точка даних приписується до найближчого кластеру.
- Після того, як усі об'єкти призначені, положення центроїдів кластерів перераховуються.
- Кроки 2 і 3 повторюються до тих пір, поки положення центроїдів не перестануть рухатися.

Визначення центральної точки залежить від вибору значення  $K$ , на яке фокусується алгоритм; це безпосередньо впливає на результати кластеризації, такі як локальна оптимальність або глобальна оптимальність. Вибір значення  $K$  безпосередньо визначає кластер даних, який необхідно об'єднати в кілька кластерів. Перевагою методу є швидкість обрахунку, відсутність вимог приналежності даних до певних груп [38].

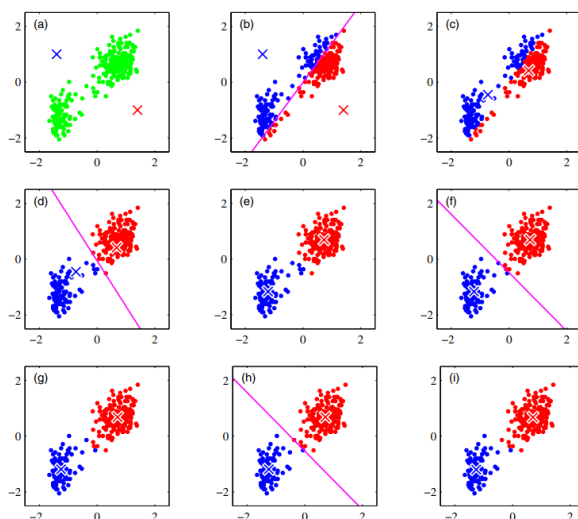


Рис. 2.4. Принцип роботи алгоритму К-середніх

Джерело: [38]

Алгоритми асоціативних правил призначені для знаходження всіх правил імплікації між умовою і наслідком. Критеріями результативності правил є рівень достовірності, підтримка, ліфт, леверидж, поліпшення. На основі застосування алгоритмів виокремлюються три види правил: корисні, тривіальні та незрозумілі. Основною перевагою асоціативних правил є легкість інтерпретації [35].

Алгоритм *Apriori* є одним із широко використовуваних алгоритмів у виявленні правил асоціацій. Він зазвичай використовується при частому видобутку набір елементів, тоді як часті набори елементів витягуються першими, а потім генеруються правила асоціації. Алгоритм використовує принцип, який стверджує, що якщо набір елементів не є частим, усі його підмножини також не є частими. Цей принцип сприяв значному покращенню часу виконання алгоритму в порівнянні з раніше використовуваним комплексним методом. Вузел *Apriori* витягує набір правил із даних, витягуючи правила з найвищим вмістом інформації. *Apriori* пропонує п'ять різних методів вибору правил і використовує схему індексування для ефективної обробки великих наборів даних. Великий набір даних підвищує швидкість тренування вибірки. *Apriori* не має довільного обмеження на кількість правил, які можна зберігати [35, 40, 41].

Алгоритм *CARMA* є іншим прикладом асоціативних правил. Для цього алгоритму попередньо не потрібно вказувати вхідних і вихідних полів призначення – можливість обмеження бажаних антецедентів і консеквентів. На першому кроці процесу подачі з'являється набір потенційних частотних наборів елементів даних. Другий крок може виконуватися не для всіх наборів транзакцій. У всьому процесі застосування алгоритм реалізує підтримку набору  $V$ , який може обробляти кожну транзакцію в  $D$  і додавати або видаляти елементи в наборі  $V$ . Коли другий обхід завершено,  $V$  є кінцевим результатом [35, 41].

*Штучна нейронна мережа* (ШНМ) складається з шарів вузлів, що містять вхідний шар, один або кілька прихованих шарів і вихідний шар. Кожен вузол або штучний нейрон з'єднується з іншим і має відповідну вагу та поріг. Якщо вихід будь-якого окремого вузла перевищує вказане порогове значення, цей вузол активується, надсилаючи дані на наступний рівень мережі. В іншому випадку дані не передаються на наступний рівень мережі. Нейронні мережі покладаються на тренувальну базу даних, щоб з часом вивчати та покращувати свою точність. Кожен окремий вузол як про власну модель лінійної регресії, що складається з вхідних даних, вагових коефіцієнтів, зміщення та вихідних даних (рисунок 2.5). Нейронна мережа може апроксимувати широкий спектр прогнозних моделей з мінімальними вимогами до структури моделі та припущень. Слабкою стороною є складність інтерпретації процесу взаємозв'язків між цільовим показником і предикатами [43, 44].

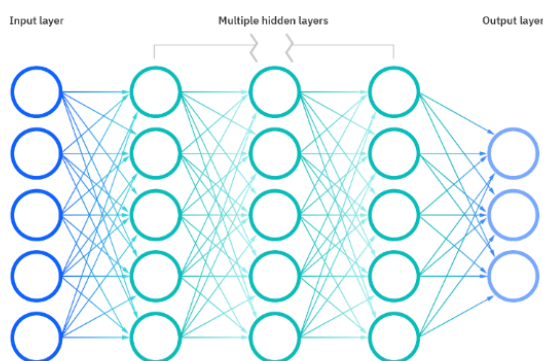


Рис. 2.5. Принцип роботи алгоритму нейронної мережі

Джерело: [44]

### 2.3. Характеристики бази даних для розробки моделей рекомендацій

Для побудови моделей в роботі використовуються три відкриті бази даних. Для реалізації рекомендаційної системи робота в IBM SPSS Modeler зосереджена на частині бази даних Amazon в розділі «Книги» виходячи з технічних можливостей реалізації робочого пристрою [45]. Зібрані дані містять інформацію 278858 учасників і 1157112 неявних та явних оцінок, посилаючись на 271379 книги.

База даних *опису книг* містить основну інформацію, згідно з якою книги ідентифікуються за відповідним ISBN. Недійсні номери ISBN уже видалено з набору даних. Крім того, надається деяка інформація на основі вмісту. Після початкової обробки база даних включає 271380 кількість записів. В таблиці 2.2 наведено наявні змінні.

Таблиця 2.2

Змінні бази даних опису книг

<i>Позначення</i>	<i>Опис</i>	<i>Тип</i>
ISBN	Ідентифікаційний код книги	Цілий
Book-Title	Назва книги	Категоріальний
Book-Author	Автор	Категоріальний
Year-Of-Publication	Рік публікації	Дата
Publisher	Видавництво	Категоріальний
Image-URL-S	Зображення в малому розмірі	Без типу
Image-URL-M	Зображення в середньому розмірі	Без типу
Image-URL-L	Зображення в великому розмірі	Без типу

Джерело: створено автором на основі [41]

База даних *рейтингів* містить інформацію про рейтинг книги. Рейтинги виражені за шкалою від 0 до 10 для 822216 записів. На основі наданих змінних розраховано додаткові змінні, що необхідні для аналізу, включаючи розрахунок оцінки продукції за власною методикою (таблиця 2.3).

Таблиця 2.3

## Змінні бази даних рейтингів

<i>Позначення</i>	<i>Опис</i>	<i>Тип</i>
User-ID	Ідентифікаційний код споживача	Цілий
ISBN	Ідентифікаційний код книги	Цілий
first_rate	Оцінка, що поставив користувач	Цілий
positive_rate	Належність до позитивних оцінок	Бінарний
negative_rate	Належність до негативних оцінок	Бінарний
number_pos_rates	Кількість позитивних оцінок в книги	Цілий
number_neg_rates	Кількість негативних оцінок в книги	Цілий
number_rates	Кількість оцінок в книги	Цілий
Year-Of-Publication	Рік публікації	Дата
number_0...10	Кількість оцінок 0..10	Цілий
Wilson	Бал Вілсона	Цілий
Bayes	Бал Байєса	Цілий
Hacker	Бал Хакер	Цілий
Own	Оцінка за власною методологією	Цілий

Джерело: створено автором на основі [41]

База даних інформації про споживачів містить дані про 278858 користувачів (таблиця 2.4). В використаній базі даних ідентифікатори користувачів (User-ID) анонімізовані та зіставлені з цілими числами, а демографічні дані надаються за наявності.

Здійснено форматування та нормалізацію баз даних та сформовано зв'язки між ними в програмі Microsoft Access (рисунок 2.6). Показник User\_ID з'єднує базу даних інформації про споживачів та базу даних рейтингів, а показник ISBN з'єднує базу даних опису книг та базу даних рейтингів. З метою полегшення користування сформовано запит, що поєднав три бази даних в одну на основі ключів (імпортується в IBM SPSS Modeler лише одна база даних).

Таблиця 2.4

## Змінні бази даних інформації про споживачів

<i>Позначення</i>	<i>Опис</i>	<i>Тип</i>
User-ID	Ідентифікаційний код споживача	Цілий
Town	Місто	Категоріальний
Region	Регіон	Категоріальний
Country	Країна	Цілий
Age	Вік	Цілий
Age_group	Інтервал віку користувача	Категоріальний

Джерело: створено автором на основі [41]



Рис. 2.6. Зв'язки між базами даних

Джерело: сформовано автором



## РОЗДІЛ 3. ПРАКТИЧНІ АСПЕКТИ РОЗРОБЛЕННЯ МОДЕЛЕЙ ПІДТРИМКИ РЕКОМЕНДАЦІЙ

### 3.1. Концептуальні основи розробки моделей підтримки рекомендацій

Моделі підтримки рекомендацій реалізовані завдяки побудові потоку інструментів в IBM SPSS Modeler, що зображені на рисунку 3.1. Моделі включають етапи від аналізу сформованої бази даних до прогнозування оцінки та формування очікувань про користувачів. В роботі використані інструменти для завантаження та аналізу бази даних (Data Audit, Type), моделювання (К-середніх, Апріорі, Регресія) та аналізу результатів (Веб, Графік, Оцінювання, Експорт). Наведене поєднання алгоритмів забезпечує цілісність та ефективність моделей, досліджених на базі даних продукції для електронної комерції.

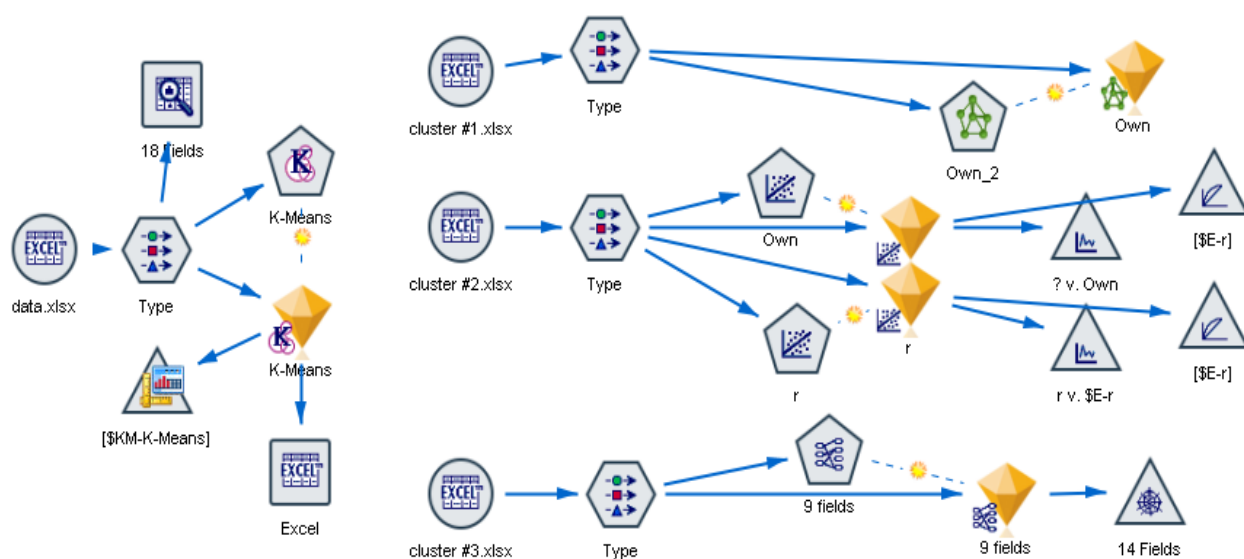


Рис. 3.1. Потік інструментів для побудови моделі в IBM SPSS Modeler

Джерело: сформовано автором

В підпункті 3.2 зосереджено увагу на порівняльній ефективності застосування розробленого методу оцінювання, що є показником бази даних. В підпункті 3.3 описано необхідність та результати застосування моделей підтримки в рекомендаційних системах.

### 3.2. Обґрунтування необхідності використання скоригованих оцінок

В параграфі описано переваги застосування скоригованого методу оцінювання продукції. Розроблено власний метод скоригованої оцінки на основі таких методів, як бал Вілсона, Байєсівська апроксимація та алгоритму ранжування Хакер (математичний опис в частині роботи 2.1). Значення, сформовано на основі методу Вілсона ( $r_w$ ) та Байєса ( $r_b$ ) рівнозначно взяті як середньоквадратична величина, скориговані на ступінь важливості ( $g$ ) та дату виробництва ( $t$ ) з методу Хакер ( $r_h$ ). Для реалізації розробленої оцінки для відібрано продукції з бази даних в середовищі R Studio написано код.

Розроблена оцінка поєднує методи на наступних принципах:

- бал методу Байєса нормалізовано до однакових меж з балом Вілсона;
- бали методів Байєса та Вілсона - середньоквадратична величина суми;
- ступінь чутливості до року виробництва 0,5;
- зведення скоригованої оцінки в 10-бальну шкалу оцінювання.

На аналізованій базі даних здійснено порівняльну характеристику оцінок за різниці методами для випадково обраних товарів (таблиця 3.1) при реалізації в середовищі R Studio.

Таблиця 3.1

Оцінки продукції за різними метриками

<i>Нормовані оцінки</i>	<i>Продукт 1</i>	<i>Продукт 2</i>	<i>Продукт 3</i>	<i>Продукт 4</i>
Вілсона	6,3218	1,3114	1,8337	0,9185
Байєса	5,4926	2,2750	1,5209	0,8726
Хакер	1,3805	3,7712	2,7925	0,2368
Середнє арифметичне	6,7632	5,3333	2,7925	1,0322
Запропонована оцінка	1,2088	1,3130	1,6845	0,2055

Джерело: сформовано автором

Продукт 1 (ISBN=439064864) має 152 оцінок серед яких 75,66% позитивні – це забезпечує високу оцінку Вілсона 6,32. Серед розподілу оцінок (таблиця 3.1)

більшість оцінок 10, 9, 8, але водночас 37 оцінок 0, тому оцінка Байєса є вище середнього значення і становить 5,49. Середнє значення 6,76. Оцінка Хакер становить лише 1,38 через те, що дата виробництва 32 роки тому. Розроблена оцінка є нижчою, ніж всі інші, оскільки враховує час публікації. Візуальний розподіл оцінок продукції 1 зображено на рисунку 3.2.

Продукція 2 (ISBN=590085417) має наступні характеристики: 3 оцінки, 2 позитивні оцінок, тяжіння до оцінки 8, застаріння виробництва 2 роки. Оцінка Вілсона 1,31, Байєса 2,27, Хакера 3,77, середнє арифметичне 5,33, скоригована власна оцінка становить 1,31 (таблиця 3.1). Розроблена модель оцінювання більшою мірою реагує як на негативні значення, так і позитивні значення (враховує ефекти різних критеріїв в синергії). Візуальний розподіл оцінок продукції 2 зображено на рисунку 3.2.

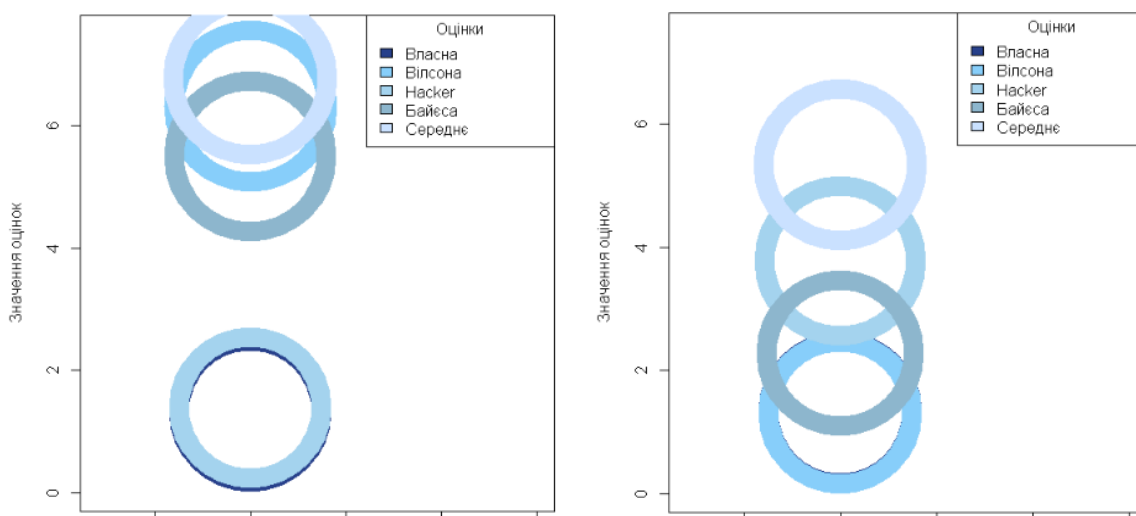


Рис. 3.2. Оцінки для продуктів 1 та 2 за різними методами оцінювання

Джерело: сформовано автором

Продукція 3 (ISBN=034545104X) має наступні характеристики: 53 оцінки, 19 позитивних оцінок, більшість оцінок 0, виробництво в останньому році з бази даних – найновіше виробництво. Значенні всіх оцінок є в межах від 1,52 до 2,79 (таблиця 3.1). Власна оцінка 1,68 є практично найнижчою серед аналізованих, що означає значний вплив хоча б однієї метрики з низьким результатом на

скориговану оцінку. Продукція 4 має низький рейтинг. Візуальний розподіл оцінок продукції 4 зображено на рисунку 3.3.

Продукція 4 (ISBN=97880107) має наступні характеристики: 2264 оцінки, 11,58% позитивних оцінок, тяжіння до оцінки 0, 18 років виробництва. Всі оцінки (таблиця 3.1) мають значення не більше 1,05 через незадовільні умови критеріїв кожного з них. Власна оцінка становить 0,21, що є нижчим значенням, ніж в інших методах. Розроблена оцінка в 5 разів менша за середнє значення, що означає вагомість кожної включеної методи до методу оцінювання. Візуальний розподіл оцінок продукції 3 зображено на рисунку 3.3.

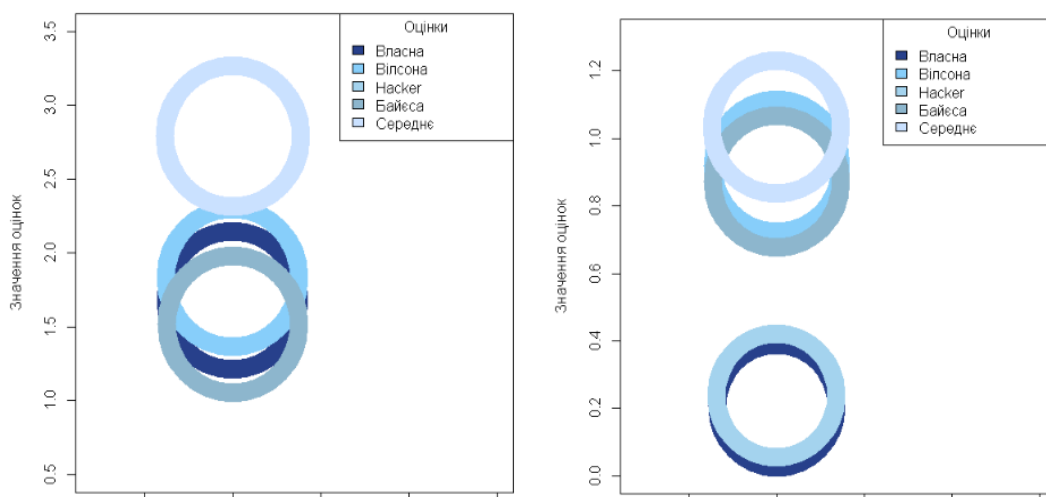


Рис. 3.3. Оцінки для продуктів 3 та 4 за різними методами оцінювання

Джерело: сформовано автором




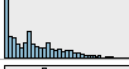
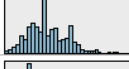
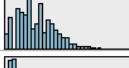
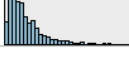
Порівняльний аналіз оцінок показав, що запропонована оцінка є більш адаптивною до використання через її чутливість до врахованих критеріїв. Вона є вищою, ніж інші, якщо всі критерії виконані на високому рівні. Водночас, оцінка є нижчою, ніж більшість інших оцінок, якщо хоча б один критеріїв є на низькому рівні. В результаті клієнт отримує до рекомендації продукції, що є більш цікавими саме для нього (персоналізованими) з врахуванням актуальності виробництва продукції, часткою позитивним відгуків. Таким чином, користувачі будуть більше купувати продукцію, що для підприємств є перевагою через збільшення збуту та виручки як головної мети діяльності.

### 3.3. Реалізація моделювання підтримки рекомендацій

В цій частині роботи покроково описано реалізацію моделей підтримки рекомендацій. Першим етапом є *завантаження бази даних та їх обробка* за допомогою інструментів Source, Data Audit. В таблиці 3.2 подано характеристику окремих показників.

Таблиця 3.2

Характеристика показників з використанням Data Audit

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
number_negative_rates		Continuous	0.000	2004.000	33.858	127.268	12.984	--	383595
number_rates		Continuous	1.000	2264.000	53.001	162.099	9.718	--	383595
date		Continuous	1901.000	2004.000	1995.310	7.424	-1.875	--	383595
Wilson		Continuous	0.000	0.716	0.126	0.113	0.835	--	383595
Bayes		Continuous	-0.056	1.230	0.313	0.158	0.501	--	383595
Hacker		Continuous	1.000	10.198	2.908	1.112	0.603	--	383595
Own		Continuous	0.001	9.831	1.234	0.989	1.785	--	383595
Country		Nominal	--	--	--	--	--	131	383595

Джерело: сформовано автором

Наступним етапом є використання методу кластеризації *K-середніх* для розбиття бази даних за групи подібних випадків. Для запуску алгоритму вказано опції для точного налаштування процесу навчання (рисунок 3.4).

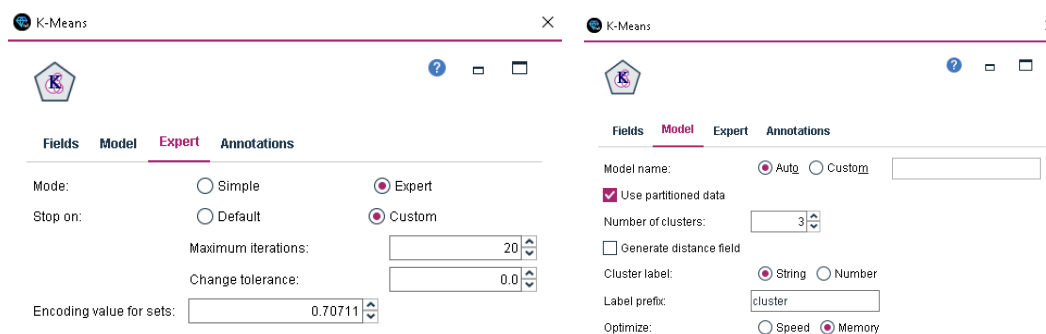


Рис. 3.4. Опції вкладок вузла К-середніх

Джерело: сформовано автором

Згідно з алгоритмом К-середніх за 15 показниками сформовано 3 кластери відмінного кількісного розподілу (рисунок 3.5, таблиця 3.3). Застосований метод має високу якість побудованих кластерів (рисунок 3.6). Найважливішими предикатами є мінімальна/максимальна оцінка, виставлена користувачем; вікова група; розроблена оцінка (рисунок 3.7).

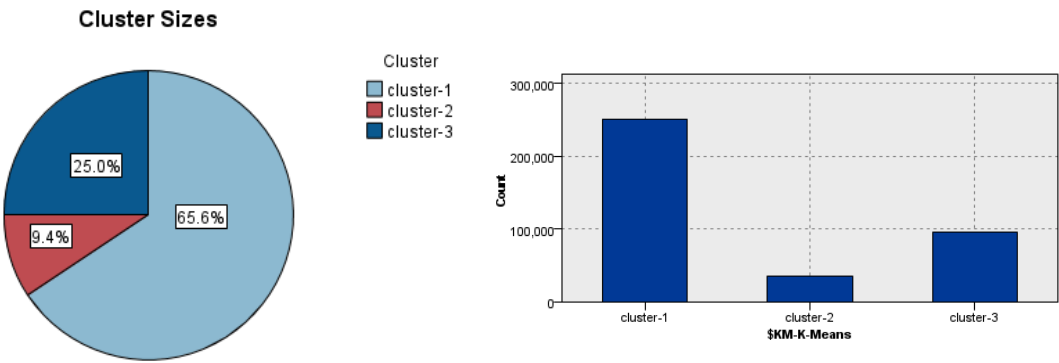
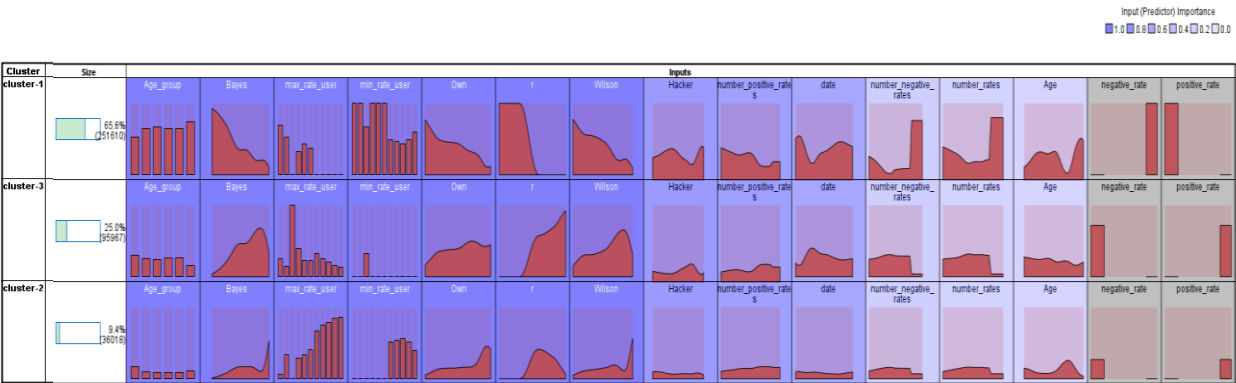


Рис. 3.5. Кількісний розподіл кластерів

Джерело: сформовано автором

Таблиця 3.3

Характеристика показників кластерів згідно з К-середніх



Джерело: сформовано автором

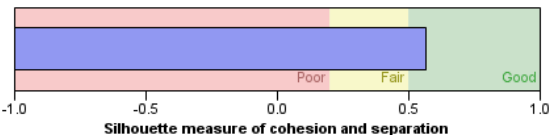


Рис. 3.6. Якість алгоритму К-середніх

Джерело: сформовано автором

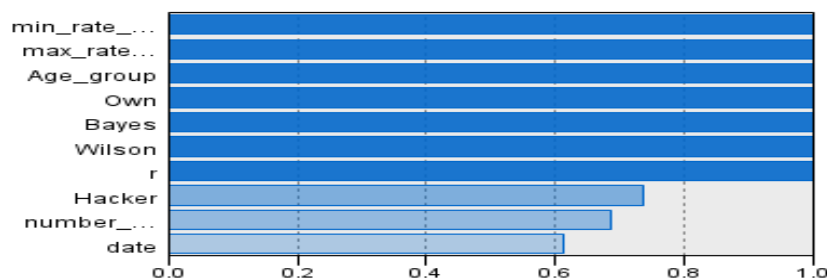


Рис. 3.7. Важливість предикатів К-середніх

Джерело: сформовано автором

На основі бази даних кожного кластера використано різні алгоритми машинного навчання, що є релевантними для рекомендаційних систем.

Для першого кластеру застосовується алгоритм *штучної нейронної мережі* (ШНМ) з вихідною змінною коригованою оцінкою продукції. Для даних цього кластера точкові значення показників були замінені на інтервальні для таких показників, як кількість позитивних/негативних оцінок продукції, рік випуску, континент користувача, кількість оцінок, виставлених користувачем, Ключові налаштування алгоритму подано на рисунку 3.8.

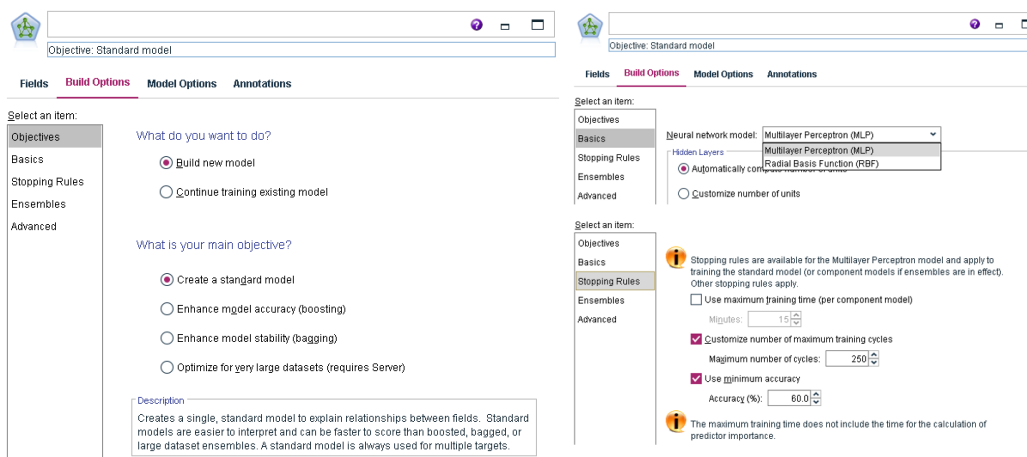


Рис. 3.8. Опції вкладок ШНМ

Джерело: сформовано автором

Побудована нейронна мережа має допустиму точність моделі загалом та важливість окремих вхідних змінних, зокрема кількості негативних та позитивних оцінок продукції; кількість оцінок, виставлених користувачем, рік

випуску продукції та інших (рисуюнок 3.9). Якість класифікації у відсотковому вимірі продемонстровано у таблиці 3.4. Вигляд нейронної мережі наочно демонструє зв'язки між змінними, що полегшує сприйняття (рисуюнок 3.10).

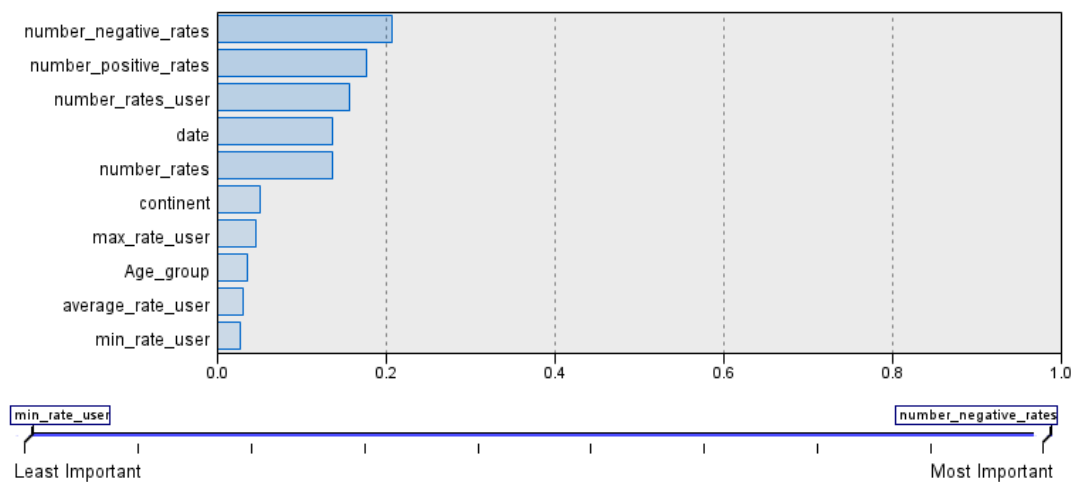


Рис. 3.9. Важливість предикатів ШНМ

Джерело: сформовано автором

Таблиця 3.4

Матриця якості класифікації

Observed	Predicted														
	0.000	1.000	2.000	3.000	4.000	5.000	6.000	7.000	8.000	9.000	10.000	11.000	12.000	13.000	15.000
0.000	0.0%	0.0%	8.7%	66.9%	24.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1.000	0.0%	0.0%	11.5%	59.4%	28.8%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
2.000	0.0%	0.0%	13.7%	54.7%	30.6%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
3.000	0.0%	0.0%	7.5%	50.1%	37.0%	4.9%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
4.000	0.0%	0.0%	3.7%	38.9%	47.3%	7.7%	2.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
5.000	0.0%	0.0%	1.9%	26.0%	40.2%	21.1%	10.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
6.000	0.0%	0.0%	1.0%	13.7%	32.7%	23.8%	25.6%	2.7%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
7.000	0.0%	0.0%	1.1%	7.2%	26.5%	17.7%	29.0%	16.0%	2.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
8.000	0.0%	0.0%	0.8%	4.4%	17.9%	5.0%	26.9%	27.5%	13.1%	0.0%	0.1%	0.0%	4.3%	0.0%	0.0%
9.000	0.0%	0.0%	1.1%	4.8%	19.0%	2.6%	24.2%	33.6%	14.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
10.000	0.0%	0.0%	0.1%	0.9%	8.5%	3.5%	2.5%	0.0%	0.0%	0.0%	84.4%	0.0%	0.0%	0.0%	0.0%
11.000	0.0%	0.0%	1.8%	9.6%	23.7%	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	0.0%	64.0%	0.0%	0.0%
12.000	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.5%	0.0%	0.0%	0.0%	97.5%	0.0%	0.0%
13.000	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
15.000	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.6%	2.6%	94.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Джерело: сформовано автором



В роботі в алгоритмі ШНМ використовувалися показники, що є доступними до відслідковування на більшості сайтах електронної комерції (включаючи ринки з невисоким рівнем розвитку рекомендаційних систем). Однак для покращення ефективності і досягнення високого рівня якості нейронної мережі необхідно апробувати ширшу кількість показників нестандартних показників різних категорій продукції перед переходом від аналізу тренувального набору даних до апробації тестового набору даних.

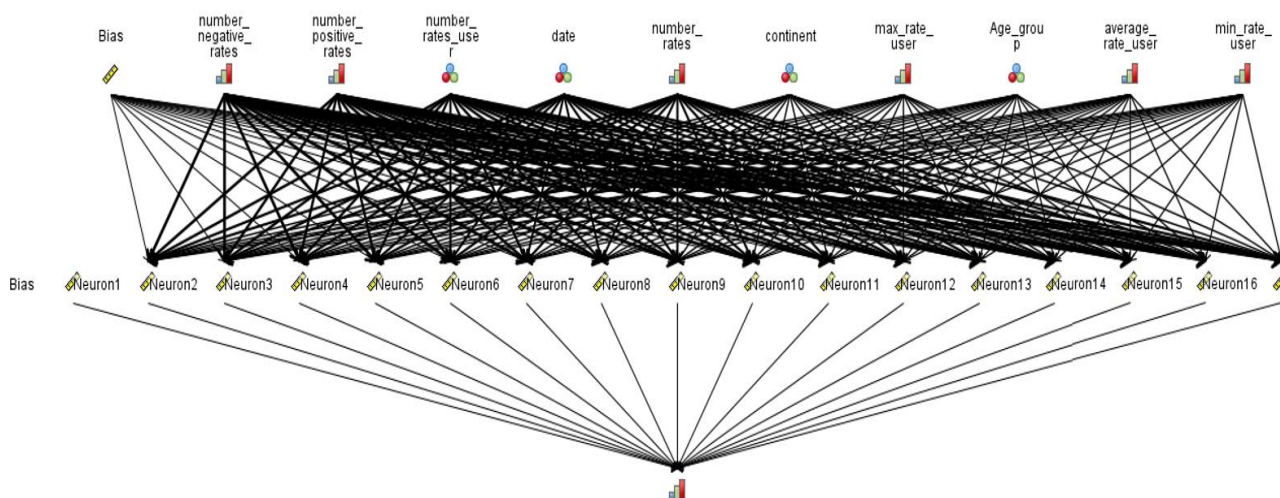


Рис. 3.10. Подання ШНМ

Джерело: сформовано автором

На основі другого кластеру застосовувалися алгоритми *лінійної регресії* для прогнозування дійсної оцінки, що виставляє користувач та розробленої методики оцінювання.

Алгоритм *лінійної регресії I* має цільову змінну – оцінку для продукції від користувача. Запуск алгоритму показав прийнятний рівень якості регресії та показників (рисунок 3.11, таблиця 3.5). Найбільш значущим показником для впливу на оцінку є кількість позитивних оцінок та дата. Алгоритм необхідний для формування та зіставлення прогнозованих очікувань і реальних типів поведінки користувачів. Результати тренувальної вибірки демонструють схожість прогнозованих та реальних значень оцінки (рисунок 3.12).

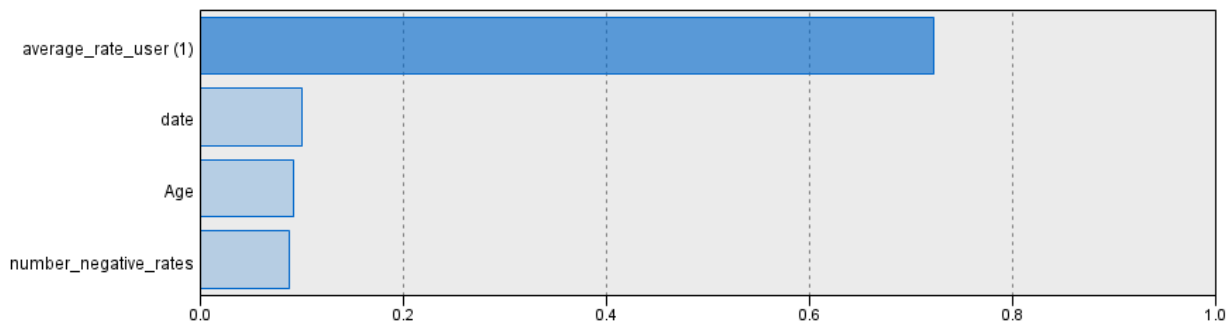


Рис. 3.11. Важливість предикатів регресії I

Джерело: сформовано автором

Таблиця 3.5

Оцінка якості регресії I та змінних

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.267 <sup>a</sup>	.071	.071	1.245615

a. Predictors: (Constant), average\_rate\_user (1), number\_negative\_rates, Age, date

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4278.705	4	1069.676	689.421	.000 <sup>b</sup>
	Residual	55750.562	35932	1.552		
	Total	60029.267	35936			

b. Predictors: (Constant), average\_rate\_user (1), number\_negative\_rates, Age, date

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.524	1.906		1.324	.185
	number_negative_rates	.000	.000	.010	1.948	.051
	date	.002	.001	.011	2.066	.039
	Age	6.891E-5	.000	.001	.140	.889
	average_rate_user (1)	.149	.003	.266	52.143	.000

Джерело: сформовано автором

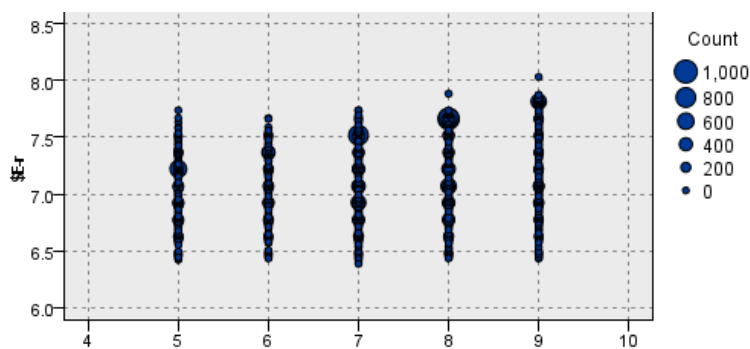


Рис. 3.12. Реальна та прогнозована оцінка продукції за методом регресії I  
Джерело: сформовано автором

Алгоритм *лінійної регресії II* з цільовою змінною скоригована оцінка за власною методикою. Модель є допустимою до використання загалом та по окремих показниках (рисунок 3.13, таблиця 3.6). Порівняння реальних та прогнозованих коригованих оцінок показують результативність алгоритму (рисунок 3.14).

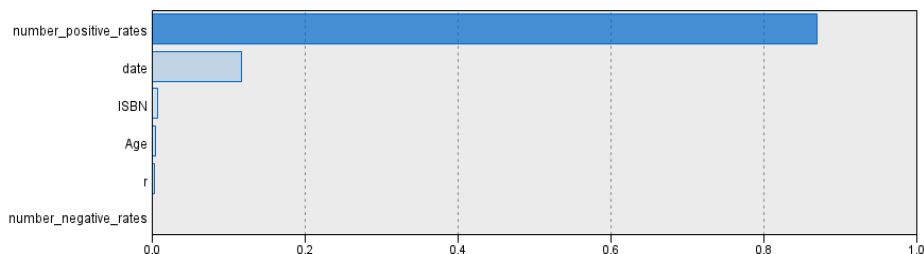


Рис. 3.13. Важливість предикатів регресії II  
Джерело: сформовано автором

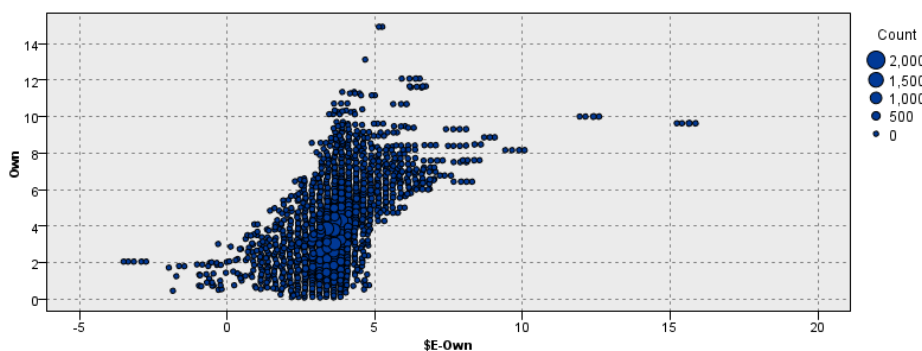


Рис. 3.14. Реальна та прогнозована оцінка продукції за методом регресії II  
Джерело: сформовано автором

Таблиця 3.6

## Оцінка якості регресії II та змінних

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.708 <sup>a</sup>	.502	.502	1.183348

a. Predictors: (Constant), ISBN, number\_negative\_rates, r, Age, date, number\_positive\_rates

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	354157.985	6	59026.331	42152.266	.000 <sup>b</sup>
	Residual	351427.939	250964	1.400		
	Total	705585.924	250970			

b. Predictors: (Constant), ISBN, number\_negative\_rates, r, Age, date, number\_positive\_rates

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-85.255	.647		-131.799	.000
	r	.144	.005	.041	29.315	<.001
	number_positive_rates	.031	.000	.835	444.742	.000
	number_negative_rates	-.004	.000	-.367	-196.794	.000
	date	.044	.000	.193	135.272	.000
	Age	-.005	.000	-.025	-17.834	<.001
	ISBN	-5.673E-14	.000	.000	-.025	.980

Джерело: сформовано автором

Для алгоритмів лінійної регресії було обраховано помилки на основі тестової бази даних (20% від початкової бази даних). Для алгоритму лінійної регресії з цільовою змінною оцінка користувача середня абсолютна помилка становить -0,8372, середньоквадратична помилка – 2,3580 і корінь середньоквадратичної помилки – 1,8051. Для алгоритму лінійної регресії з цільовою змінною скоригована оцінка середня абсолютна помилка становить 0,1203, середньоквадратична помилка - 1,4605 і корінь середньоквадратичної помилки – 0,5672. Обраховані показники помилок є допустимого рівня.

На основі бази даних третього кластеру застосовано алгоритм асоціативних правил *Apriori* з визначенням основних критеріїв підтримки виконання правил (рисунок 3.15) у наборі даних. Алгоритм враховує 9 консеквенти та 2 антецеденти (рисунок 3.16). У результаті роботи алгоритму одержано 78 правил, частина яких подано в таблиці 3.7. Визначені правила виокремлюють найбільш часті випадки залежності показників, що впливають на оцінювання та мають застосування для прогнозування. Для оцінки зв'язку між категоріями продукції використано вузол Web, що зображено на рисунку 3.17.

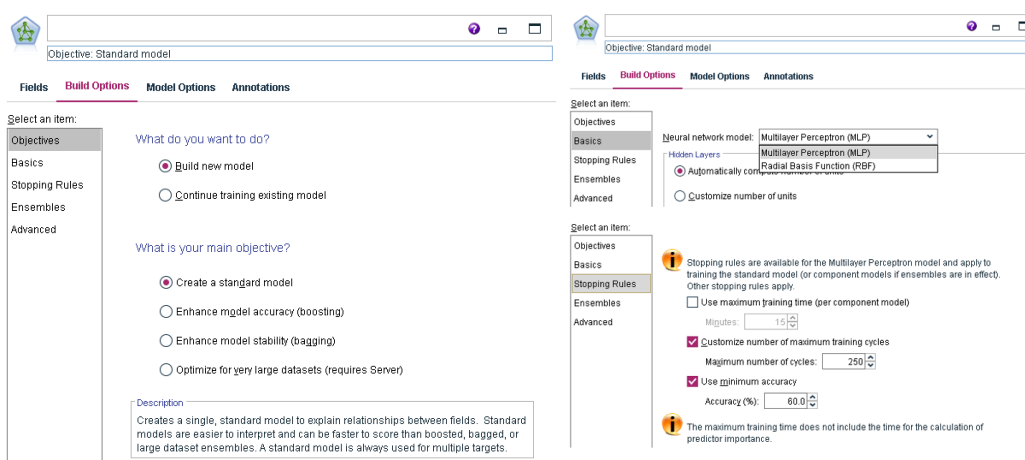


Рис. 3.15. Опції вкладок Апріорі

Джерело: сформовано автором

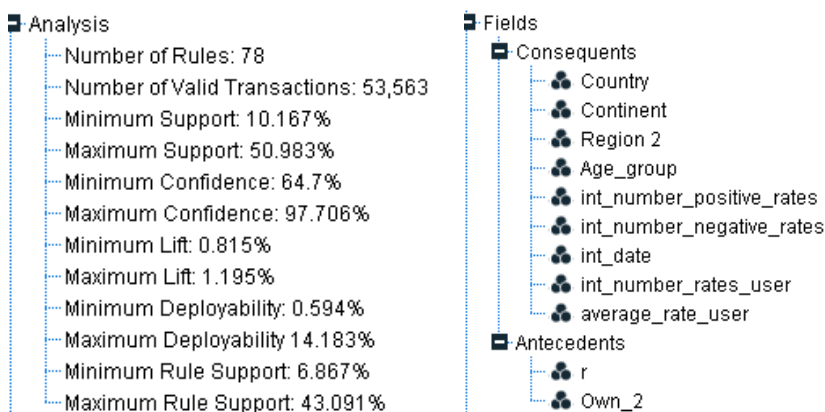


Рис. 3.16. Показники алгоритму Апріорі та результати підтримки правил

Джерело: сформовано автором

Таблиця 3.7

## Оцінка якості регресії II та змінних

Консеквент	Антецедент	Екземпляри	Підтримка	Впевненість	Підтримка правила	Ліфт
Region 2 = Northern America	r = 10.0 and Own_2 = 1.0	6526	12,18	85,67	10,44	1,02
int_number_negative_rates = <15	r = 8.0 and Own_2 = 1.0	6843	12,78	84,39	10,78	1,13
int_date = <2004	average_rate_user = 5.0	5446	10,17	84,32	8,57	1,03
Continent = North America	r = 8.0 and Own_2 = 1.0	6843	12,78	84,26	10,76	1,00
int_date = <2004	average_rate_user = 2.0	9689	18,09	82,2	14,87	1,01
Continent = North America	r = 7.0	8386	15,66	80,97	12,68	0,96
Country = usa	average_rate_user = 1.0	7578	14,15	80,25	11,35	1,08
Region 2 = Northern America	average_rate_user = 3.0	9412	17,57	79,92	14,04	0,95
int_date = <2004	r = 10.0	13320	24,87	79,41	19,75	0,97
Continent = North America	average_rate_user = 5.0	5446	10,17	78,28	7,96	0,93
Region 2 = Northern America	average_rate_user = 4.0	6342	11,84	78,16	9,25	0,93
Region 2 = Northern America	average_rate_user = 5.0	5446	10,17	78,04	7,93	0,93
Country = usa	r = 10.0 and Own_2 = 1.0	6526	12,18	76,52	9,32	1,03
int_number_positive_rates = <15	r = 8.0 and Own_2 = 1.0	6843	12,78	74,73	9,55	1,00
int_date = <2004	r = 8.0 and Own_2 = 1.0	6843	12,78	72,82	9,30	0,89
int_date = <2004	Own_2 = 1.0	27308	50,98	72,18	36,8	0,88
int_number_negative_rates = <15	average_rate_user = 2.0	9689	18,09	72,09	13,04	0,96
int_number_negative_rates = <15	Own_2 = 2.0	13864	25,88	64,70	16,75	0,86

Джерело: сформовано автором

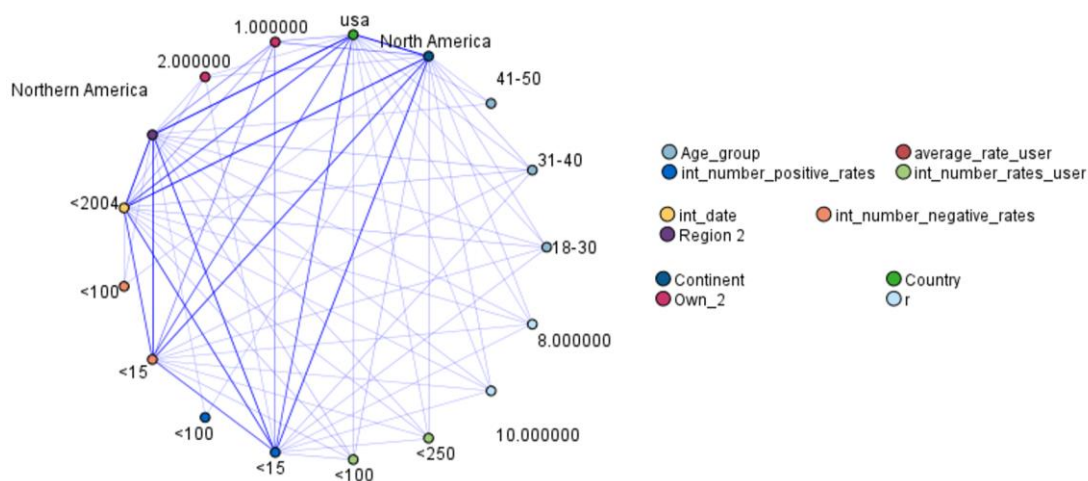


Рис. 3.17. Графік частоти зв'язку між показниками

Джерело: сформовано автором

Моделі підтримки рекомендацій на основі використання таких алгоритмів машинного навчання, як К-середніх, Априорі, регресійна залежність, асоціативні правила урізноманітнюють способи використання найбільш часто використовуваних даних з підвищенням ефективності моделей разом у сукупності. Поділ усієї бази даних на кластери забезпечує визначення особливостей випадків користувачів та продукції. В роботі для бази даних кожного кластеру наведено по одному типу алгоритму машинного навчання. Для практичної реалізації необхідно ж здійснювати порівняльний аналіз ефективності різних алгоритмів для окремих кластерів, враховуючи їх характеристики.

## ВИСНОВКИ

Результатом кваліфікаційної роботи бакалавра з економічної кібернетики є теоретичне узагальнення та вирішення науково-практичного завдання, що полягає в удосконаленні концепції та реалізації моделей підтримки рекомендацій в сфері електронної комерції. Проведене дослідження дозволяє сформулювати такі висновки і пропозиції:

1. Існує різноманітність рекомендаційних систем залежно від типів та методів, проте кожна з них є експертною системою фільтрації інформації, метою функціонування якої є передбачення потенційних потреб користувачів шляхом аналізу їх уподобань та надання персоналізованих рекомендаційних послуг.

2. Приклади удосконалень гібридних рекомендаційних систем зосереджені на подоланні проблем холодного старту, довгому хвості продукції, тональності даних, урахуванні телеметрії, поєднанні таксономії та фольклонії, урахуванні коефіцієнтів подібності профілів та демографічних відмінностей, співставленні опису продукції і профілю користувача. Підвид сучасних рекомендаційних систем застосовує алгоритми машинного навчання для вирішення вищезгаданих проблем.

3. Для ринку з невисоким рівнем розвитку рекомендаційних технологій (зокрема і українського) характерні певні проблеми, а саме використання лише прямої інформації користувача або загальної інформації усіх користувачів в сукупності, обмеженість використовуваних алгоритмів рекомендаційних систем, обрахунок оцінок продукції за найпростішими метриками. Їх частково можна вирішити, використовуючи специфічні моделі систем і специфічні методи оцінювання, представлені в дослідженні.

4. Запропонована модель оцінювання продукції дає можливість одержати більш точну оцінку продукції. Це дозволяє сформувати для користувачів більш персоналізовані рекомендації, що якісніше задовольняє їх попит. Для підприємств це забезпечує зростання іміджу бренду у зв'язку з тим, що компанія детальніше вивчає кожного споживача, а також забезпечує



збільшення збуту продукції та приріст виручки відповідно. В рекомендаційній системі оцінювання відбувається на основі поєднанні методів Вілсона, Байєса та Хакер при їх нормалізації.

5. Аналіз результативності запропонованої метрики в рекомендаційній системі у порівнянні з базовими методиками показав, що розроблена оцінка є точнішою через її чутливість до врахованих критеріїв. Вона є вищою, ніж інші, якщо всі критерії виконані на високому рівні. Водночас, оцінка є нижчою, ніж більшість інших оцінок, якщо хоча б один критеріїв є на низькому рівні. Тому її варто застосовувати, коли представлена продукція на сайті електронної комерції має широкі розмахи за такими показниками, як час, кількість оцінок, розподіл позитивних і негативних оцінок тощо.

6. Розроблені моделі підтримки рекомендацій застосовувалися для бази даних електронної комерції. Для даних категорії «Книги» оцінка продукції скоригована за розробленим методом оцінювання. Початкова база даних розподілена на кластери за допомогою алгоритму К-середніх. На основі кластерів використовуються алгоритми підтримки рекомендацій, такі як нейронна мережа, лінійна регресія, асоціативні правила, що реалізовані в аналітичній платформі IBM SPSS Modeler. Результати навчання моделей демонструють допустимий рівень якості. Моделі є ефективними та придатними до використання завдяки різнобічному підходу до аналізу простих та найбільш використовуваних метрик з сайтів електронної комерції.

7. Напрямки удосконалення запропонованих моделей підтримки рекомендацій для можливості їх реалізації на різних платформах електронної комерції передбачають апробацію на базі даних вітчизняних компаній з широким вибором категорій продукції та здійсненні порівняльного аналізу алгоритмів машинного навчання з врахуванням особливостей ринків з невисоких рівнем розвитку рекомендаційних технологій.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Feng J., Xia Z., Feng X., Peng J. RBPR: A hybrid model for the new user cold start problem in recommender systems. Knowledge-Based Systems. 2021. 214. doi: 10.1016/j.knosys.2020.106732
2. Лі К. В. Дослідження та розробка гібридної рекомендаційної системи на прикладі музичного веб-сервісу : атестаційна робота здобувача вищої освіти на другому (магістерському) рівні. Харків, 2019. 62 с.
3. Apáthy S. History of recommender systems. Onespire. URL: <https://www.onespire.net/news/history-of-recommender-systems/> (date of access: 02.06.2022)
4. Qomariyah N. Definition and History of Recommender Systems. Binus University International. 2020. URL: <https://international.binus.ac.id/computer-science/2020/11/03/definition-and-history-of-recommender-systems/>
5. Шварц М. Є. Гібридні моделі і методи прогнозування рекомендацій для інтернет-магазину: дисертаційна робота. Львів, 2019. 152 с.
6. Bhanuse, R., Mal, S. A Systematic Review: Deep Learning based E-Learning Recommendation System. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). 2021. P. 190-197. doi:10.1109/icaais50930.2021.9395835
7. Woldan P., Duda P., Hayashi Y. Visual Hybrid Recommendation Systems Based on the Content-Based Filtering. 19th International Conference on Artificial Intelligence and Soft Computing. 2020. 12416. P. 455-465. doi: 10.1007/978-3-030-61534-5\_41
8. Barolli L., Cicco F., Fonisto M. An Investigation of Covid-19 Papers for a Content-Based Recommendation System. International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. 2021. 343. P. 156-164. doi:10.1007/978-3-030-89899-1\_16

9. Dat N., Toan P., Thanh T. Solving distribution problems in content-based recommendation system with gaussian mixture model. *Applied Intelligence*. 2021. P. 1062 - 1614. doi: 10.1007/s10489-021-02429-9
10. Kumar B., Kumar P. Fattening The Long Tail Items in E-Commerce. *Journal of Theoretical and Applied Electronic Commerce Research* . 2017. 12(3). P. 27-49. doi: 10.4067/S0718-18762017000300004
11. Ajaegbu C. An optimized item-based collaborative filtering algorithm. *Journal of ambient intelligence and humanized computing*. 2021. 12. P. 10629-10636. doi <https://doi.org/10.1007/s12652-020-02876-1>
12. Nhuen L., Hong M., Jung J., Sohn B. Cognitive Similarity-Based Collaborative Filtering Recommendation System. *Applied Sciences*. 2020. 10 (12). doi: 10.3390/app10124183
13. Shen J., Zhou T., Chen L. Collaborative filtering-based recommendation system for big data. *International Journal of Computational Science and Engineering*. 2020. 21(2). P. 219-225. doi: 10.1504/IJCSE.2020.105727
14. Chornous G., Nikolskyi I., Wyszyński M., Kharlamova G., Stolarczyk P. A hybrid user-item-based collaborative filtering model for e-commerce recommendations. *Journal of International Studies*. 2021. 14(4). P. 157-173. doi: 10.14254/2071-8330.2021/14-4/11
15. Чорна К. Ю., Замятін Д. С. Система рекомендацій з використанням соціальних мереж: магістерська робота. Київ, 2018. 71 с.
16. Y. Kilani, A. F. Otoom, A. Alsarhan, M. Almaay, A genetic algorithms-based hybrid recommender system of matrix factorization and neighborhood-based techniques. *Journal of Computational Science*. 2018. 8. P. 78-93. doi: 10.1016/j.jocs.2018.08.007
17. Jiang. Hybrid Recommendation Approaches. iCAMP. UCI Interdisciplinary Computational and Applied Mathematics Program. URL: [https://www.math.uci.edu/~icamp/courses/math77b/lecture\\_12w/](https://www.math.uci.edu/~icamp/courses/math77b/lecture_12w/) (data of access 02.06.2022)

18. Rudolf Turnip, Dade Nurjanah, Dana Sulistyo Kusumo. Hybrid recommender system for learning material using content-based filtering and collaborative filtering with good learners' rating. IEEE Conference on e-Learning, e-Management and e-Services (IC3e). 2017. P. 61-68. doi: 10.1109/IC3e.2017.8409239
19. Lopatka M., Ng V., Miroglio, B.P., (...), Placitelli A., Thomson L. Telemetry-Aware Add-on Recommendation for Web browser customization. ACM UMAP 2019. 27th ACM Conference on User Modeling, Adaptation and Personalization. 2019. P.166-175. doi: 10.1145/3320435.3320450
20. Mao M., Chen S., Zhang F., Han J., Xiao Q. Hybrid ecommerce recommendation model incorporating product taxonomy and folksonomy. Knowledge-Based Systems. 2021. 214. doi: 10.1016/j.knosys.2020.106720
21. Patil M., Rao M. Studying the Contribution of Machine Learning and Artificial Intelligence in the Interface Design of E-commerce. Smart Intelligent Computing and Applications. 2018. 105 (2). P. 197-206. doi: 10.1007/978-981-13-1927-3\_20
22. Zhou L. Product advertising recommendation in e-commerce based on deep learning and distributed expression. Electron Commer Res 20. 2020. P. 321–342. doi: 10.1007/s10660-020-09411-6
23. Kumar S., Anthoniraj A. Probabilistic Unsupervised Machine Learning Approach for a Similar Image Recommender System for E-Commerce. Symmetry. 2020. 12 (11). doi: 10.3390/sym12111783
24. Anitha J., Kalaiarasu M. Optimized machine learning based collaborative filtering (OMLCF) recommendation system in e-commerce. Journal of Ambient Intelligence and Humanized Computing. 2021. 12(4). P. 6387–6398. doi: doi.org/10.1007/s12652-020-02234-1
25. Addagarla K., Amalanathan A. Probabilistic Unsupervised Machine Learning Approach for a Similar Image Recommender System for E-Commerce. Symmetry. 2020. 12(11). doi: 10.3390/sym12111783
26. Singh M., Rishi O. Event driven Recommendation System for E-commerce using Knowledge based Collaborative Filtering Technique. Scalable

Computing. Practice and Experience. 2020. 21. P. 369-378. doi: 10.12694/scpe.v21i3.1709.

27. Zhao Q., Zhang Y., Friedman D., Fangfang T.. E-commerce Recommendation with Personalized Promotion. 9<sup>th</sup> ACM Conference on Recommender Systems. 2015. P. 219-226. doi: 10.1145/2792838.2800178.

28. Miller E. How to Sort By Average Rating. Evanmiller. 2009. URL: <https://www.evanmiller.org/how-not-to-sort-by-average-rating.html> (date of access: 02.06.2022)

29. Kumar A. Wilson Lower bound Score and Bayesian Approximation for K star scale rating to Rate products. Tech-that-works. 2020. URL: <https://medium.com/tech-that-works/wilson-lower-bound-score-and-bayesian-approximation-for-k-star-scale-rating-to-rate-products-c67ec6e30060>

30. Sallihendic A. How Hacker News ranking algorithm works. Tech-that-works. 2015. URL: <https://medium.com/hacking-and-gonzo/how-hacker-news-ranking-algorithm-works-1d9b0cf2c08d>

31. Ziegler C., McNee S., Konstan J., Lausen G. Improving recommendation lists through topic diversification. 14th International World Wide Web Conference (WWW '05). 2005. P. 22-35. doi: 10.1145/1060745.1060754

32. Nehrey M., Hnot T. Using recommendation approaches for ratings matrixes in online marketing. Studia Ekonomiczne. 2017. 342. P. 115-130.

33. Chen M., Liu P. Performance Evaluation of Recommender Systems. Int J Performability Eng. 2017. 3(8). P. 1246-1256. doi: 10.23940/ijpe.17.08.p7.12461256

34. Lendare V. How to Measure the Success of a Recommendation Systems? Developers corner. 2021. URL: <https://analyticsindiamag.com/how-to-measure-the-success-of-a-recommendation-system/>

35. Черняк О. І., Чорноус Г. О. Інтелектуальний аналіз даних у бізнесі з використанням IBM SPSS Modeler: навчальний посібник. ВПЦ «Київський університет». 2020. – 263 с.

36. Nawrocka A., Kot A., Nawrocki M. Application of machine learning in recommendation systems. 19th International Carpathian Control Conference (ICCC). 2018. P. 328-331. doi: 10.1109/CarpathianCC.2018.8399650.
37. K-Means Clustering. Techopedia. URL: <https://www.techopedia.com/definition/32057/k-means-clustering> (date of access: 02.06.2022)
38. Yuan C., Yang H. Research on K-Value Selection Method of K-Means Clustering Algorithm. J. 2019. 2(2). P. 226-235. doi: 10.3390/j2020016
39. Bishop M. Pattern Recognition and Machine Learning. Information Science and Statistics. Edited by M Jordan, J Kleinberg and B Schölkopf. 2006. 4(4).
40. AlZu'bi S., Hawashin B., ElBes M., Al-Ayyoub M. A Novel Recommender System Based on Apriori Algorithm for Requirements Engineering. Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2018. P. 323-327. doi: 10.1109/SNAMS.2018.8554909.
41. Association Rules node. IBM. URL: <https://www.ibm.com/docs/en/cloud-paks/cp-data/4.0?topic=modeling-association-rules-node> (date of access: 02.06.2022)
42. Hu L. Research on English Achievement Analysis Based on Improved CARMA Algorithm. Advanced Computational Intelligence Algorithms for Signal and Image Processing. 2022. doi: 10.1155/2022/8687879
43. Neural Net node. IBM. URL: <https://www.ibm.com/docs/en/cloud-paks/cp-data/4.0?topic=modeling-neural-net-node> (date of access: 02.06.2022)
44. Neural Networks. IBM. URL: <https://www.ibm.com/cloud/learn/neural-networks> (date of access: 02.06.2022)
45. Ziegler C., McNee M., Konstan J., Lausen G. Improving recommendation lists through topic diversification. Proceedings of the 14th International World Wide Web Conference (WWW '05). 2005. P. 25-32. doi: 10.1145/1060745.1060754