

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма	«Економічна кібернетика»
Галузь знань	05 «Соціальні та поведінкові науки»
Спеціальність	051 «Економіка»

Форма навчання: очна (денна)
очна (денна) / дистанційна

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему « **Аналіз та моделювання рейтингу фільмів (на прикладі
рейтингу IMDb)**»
(назва теми)

здобувача Константинова Михайла Валерійовича
(ПІБ)

(підпис)

Науковий керівник: кандидат фізико-математичних наук,
професор Великоіваненко Г.І.

(підпис)

**Робота допущена до захисту перед екзаменаційною
комісією з атестації здобувачів вищої освіти (ЕК)**

В.о. завідувача кафедри: кандидат фізико-математичних наук,
професор Великоіваненко Г.І.

(підпис)

Київ 2024

ЗМІСТ

ВСТУП	3
РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ РЕЙТИНГУВАННЯ ТА ОСОБЛИВОСТІ ЙОГО ЗАСТОСУВАННЯ У КІНОІНДУСТРІЇ.....	8
1.1 Сутність рейтингового управління в економіці	8
1.2 Особливості кіноіндустрії та необхідність рейтингування фільмів.....	14
1.3 Особливості IMDb та чинники формування його рейтингу.....	20
РОЗДІЛ 2. МОДЕЛЮВАННЯ РЕЙТИНГУ ФІЛЬМІВ НА ПРИКЛАДІ РЕЙТИНГУ IMDB.....	27
2.1 Моделювання рейтингу на підґрунті кластерного аналізу.....	27
2.2 Моделювання рейтингу на підґрунті карт Кохонена	36
2.3 Порівняльний аналіз отриманих результатів та практичні рекомендації.....	47
ВИСНОВКИ.....	55
ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	59
ДОДАТКИ... ..	62

ВСТУП

Кіноіндустрія займає важливе місце в сучасному культурному та економічному житті суспільства. Вона не лише формує культурний контекст та впливає на світогляд мільйонів людей, але й є ваговою складовою глобальної економіки. Фільми відображають соціальні, політичні та культурні тенденції, створюючи унікальний історичний запис та впливаючи на громадську думку. Кіно також виступає потужним засобом освіти і просвіти, дозволяючи аудиторії розширювати свої горизонти та отримувати нові знання про світ. Окрім цього, кіноіндустрія є важливим джерелом розваг, що об'єднує людей різного віку, національностей та культур. Завдяки технологічним інноваціям та розширенню доступу до кінофільмів через інтернет, кіноіндустрія постійно розвивається, забезпечуючи робочі місця для мільйонів людей по всьому світу.

Зі збільшенням кількості стрімінгових сервісів та цифрових платформ, доступ до фільмів став значно простішим, що лише підсилює важливість кіноіндустрії у сучасному суспільстві. Одним з ключових елементів успішного функціонування кіноіндустрії є система рейтингування фільмів. Рейтингування дозволяє оцінювати якість кінопродукції, допомагає глядачам у виборі фільмів для перегляду та орієнтує їх у світі кіно. Завдяки системі рейтингування глядачі можуть робити обґрунтовані вибори, обираючи ті фільми, які найбільше відповідають їхнім уподобанням та очікуванням. Рейтингування також стимулює виробників фільмів до створення якісного контенту, оскільки високі рейтинги можуть значно вплинути на касові збори та загальний успіх фільму.

Дипломна робота присвячена аналізу та моделюванню рейтингів фільмів на прикладі рейтингу IMDb. Метою даної роботи є дослідження особливостей рейтингування фільмів (на прикладі рейтингу IMDb), аналіз факторів, які впливають на його формування, та вибір адекватного інструментарію моделювання рейтингу фільмів IMDb.

У рамках дослідження буде проведено аналіз чинників, що впливають на формування рейтингів, а також розглянуто методи кластерного аналізу та самоорганізуючі карти Кохонена (SOM) як інструменти для моделювання рейтингів.

Основні завдання роботи включають:

- Аналіз сутності рейтингового управління в економіці з урахуванням його особливостей та впливу на зовнішні чинники.
- Вивчення особливостей кіноіндустрії та обґрунтування необхідності рейтингування фільмів в цьому контексті.
- Аналіз платформи IMDb як найбільш відомого та використовуваного ресурсу для рейтингування фільмів, включаючи огляд чинників, що впливають на формування рейтингу на цій платформі.
- Розробка та порівняльний аналіз моделей моделювання рейтингу фільмів на прикладі IMDb з використанням кластерного аналізу та карт Кохонена.
- Формулювання висновків на основі результатів дослідження та надання практичних рекомендацій щодо використання моделей рейтингування в практичній діяльності в кіноіндустрії.

Об'єктом дослідження є процес рейтингування фільмів, що є складним та багатограним явищем, яке охоплює збір, аналіз та інтерпретацію великої кількості даних. Ці дані відображають думки та оцінки глядачів, критиків та інших учасників ринку. Основна проблема цього процесу полягає у необхідності об'єктивного та адекватного оцінювання якості та популярності фільмів. Це, в свою чергу, впливає на прийняття рішень у кіновиробництві, маркетингу та розповсюдженні фільмів.

Предметом дослідження є методи та моделі аналізу і рейтингування фільмів на прикладі рейтингу IMDb, зокрема використання кластерного аналізу та карт Кохонена для моделювання та прогнозування рейтингів фільмів та конкретизація проблемної ситуації рейтингування.

Для досягнення мети кваліфікаційної бакалаврської роботи буде використано сучасні методи дослідження, зокрема кластерний аналіз та самоорганізуючі карти Кохонена. Кластерний аналіз дозволяє групувати фільми за

схожими характеристиками та визначати основні чинники, які впливають на їхній рейтинг. Самоорганізуючі карти Кохонена є потужним інструментом для візуалізації багатовимірних даних, що дозволяє легко інтерпретувати результати аналізу та знаходити приховані закономірності.

Інформаційна база дослідження складається з різноманітних джерел даних та інформації, а саме: база даних IMDb, наукові статті, статистичні звіти та аналітичні огляди, інтернет-ресурси та бази-даних.

Рейтингування фільмів, зокрема на таких платформах як IMDb (Internet Movie Database), надає важливу інформацію як для кінолюбителів, так і для професіоналів кіноіндустрії. IMDb є однією з найвідоміших та найвпливовіших платформ, що забезпечує доступ до величезного обсягу інформації про фільми, серіали, акторів, режисерів та інші аспекти кіноіндустрії. IMDb пропонує користувачам можливість залишати свої оцінки та рецензії на фільми, що створює великий масив даних, який використовується для формування рейтингів. Основним елементом платформи є рейтинги фільмів, які формуються на основі оцінок користувачів. Ці рейтинги відіграють значну роль у прийнятті рішень глядачами щодо перегляду певного фільму, а також впливають на комерційний успіх кінострічок. Високий рейтинг на IMDb може привернути більше глядачів, збільшити продаж квитків та сприяти успіху фільму на міжнародному ринку. Крім того, рейтинги фільмів на IMDb часто використовуються критиками та кіномитцями як орієнтир для оцінки якості кінопродукції. В умовах зростаючої конкуренції в кіноіндустрії, розуміння процесів формування рейтингів та їх впливу на популярність фільмів стає все більш актуальним. Платформи на кшталт IMDb не лише впливають на індивідуальні рішення глядачів, але й формують загальні тенденції та вподобання в кіно. Аналіз процесів формування рейтингів дозволяє виявити основні чинники, які впливають на оцінки користувачів, та зрозуміти, як можна покращити якість та привабливість кінопродукції. Це знання є цінним як для кіновиробників, так і для маркетологів, які прагнуть ефективно просувати фільми та залучати широку аудиторію. Таким чином, рейтингування фільмів є невід'ємною частиною сучасної кіноіндустрії, яка допомагає глядачам орієнтуватися у

величезному масиві кінопродукції, а виробникам фільмів – підвищувати якість своєї роботи та досягати комерційного успіху.

Система рейтингування фільмів відіграє ключову роль у кіноіндустрії, оскільки вона впливає на рішення глядачів щодо перегляду того чи іншого фільму. Вона також має значний вплив на маркетингові стратегії та фінансові результати кіностудій. З огляду на це, важливо розуміти, як формуються ці рейтинги і які чинники на них впливають. Сучасні технології дозволяють аналізувати великі обсяги даних, що відкриває нові можливості для удосконалення процесів рейтингування.

Одним з найбільш впливових та популярних джерел рейтингів фільмів є платформа IMDb, яка збирає та обробляє оцінки мільйонів користувачів по всьому світу. Рейтинг фільму на IMDb формується на основі середньозважених оцінок користувачів, що дозволяє отримати об'єктивну картину популярності та якості фільму. Однак, сам процес формування рейтингу є складним та багатофакторним, що вимагає глибокого аналізу та розуміння.

Розвиток технологій та зростання обсягів даних відкриває нові можливості для застосування складних та ефективних методів аналізу і моделювання інформації. В цьому контексті особливої уваги заслуговують методи кластерного аналізу та самоорганізуючих карт, які дозволяють структурувати та візуалізувати великі масиви даних, знаходячи в них приховані закономірності. Кластерний аналіз дозволяє виявляти групи об'єктів, що мають схожі характеристики, що є важливим для сегментації ринку та цільового маркетингу. Самоорганізуючі карти Кохонена забезпечують потужний інструмент для багатовимірної візуалізації, що полегшує їх аналіз та інтерпретацію.

Важливим аспектом роботи є не лише аналіз існуючих методів рейтингування, але й розробка нових підходів до моделювання рейтингів фільмів. Це дозволить не лише покращити точність та об'єктивність рейтингів, але й виявити нові фактори, що впливають на оцінки користувачів. В рамках роботи будуть розглянуті різні методи та підходи до моделювання, а також проведено порівняльний аналіз їхньої ефективності.

Окрім наукової цінності, результати дослідження мають значний практичний потенціал. Використання методів кластерного аналізу та самоорганізуючих карт дозволяє не лише підвищити точність прогнозування рейтингів, але й сприяти кращому розумінню структури та динаміки аудиторії кінофільмів. Це може бути корисним інструментом для багатьох фахівців кіноіндустрії, що прагнуть оптимізувати свої стратегії просування та підвищити конкурентоспроможність своєї продукції.

Також варто зазначити, що впровадження нових технологій та підходів до аналізу даних сприятиме подальшому розвитку кіноіндустрії в цілому. Інноваційні методи аналізу дозволяють отримати більш глибоке та точне розуміння вподобань аудиторії, що відкриває нові можливості для створення контенту, який відповідає потребам і очікуванням глядачів. Це, в свою чергу, сприяє підвищенню задоволеності глядачів та збільшенню прибутків від кінофільмів.

Отже, дана робота має на меті всебічне вивчення процесу рейтингування фільмів та розробку ефективних методів моделювання рейтингів, що сприятиме підвищенню якості оцінювання кінопродукції та допоможе глядачам зробити більш обґрунтований вибір при виборі фільмів для перегляду. Робота буде корисною не лише для науковців та дослідників, але й для практиків кіноіндустрії, які займаються маркетингом, просуванням та оцінюванням кінопродукції. Завдяки використанню сучасних методів аналізу даних та моделювання, ця робота має на меті внести вагомий внесок у розвиток теорії та практики рейтингування фільмів, а також допомогти у вирішенні актуальних проблем кіноіндустрії.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ОСНОВИ РЕЙТИНГУВАННЯ ТА ОСОБЛИВОСТІ ЙОГО ЗАСТОСУВАННЯ У КІНОІНДУСТРІЇ

1.1 Сутність рейтингового управління в економіці

Рейтингове управління — це складний процес, який включає в себе використання рейтингів як інструментів для аналізу, оцінки та ухвалення рішень у різноманітних сферах діяльності. Ці рейтинги можуть базуватися на різних критеріях і методологіях, залежно від специфіки об'єкта оцінки та цілей, які переслідуються. Ключові аспекти рейтингового менеджменту включають створення рейтингів, аналіз та використання рейтингів для цілей управління [1].

Формування рейтингів — це процес визначення критеріїв, за якими буде оцінюватися об'єкт, та збір інформації для цієї оцінки. Критерії можуть бути кількісними (наприклад, фінансові показники) або якісними (наприклад, рівень задоволеності клієнтів). Цей процес також включає вибір методології оцінки, яка може бути заснована на статистичних моделях, експертних оцінках або комбінації обох підходів. Наприклад, у роботі [2] автори досліджують процеси формування та класифікації методів рейтингів. Вони наголошують на важливості науково-методологічного підходу до розробки рейтингів та їх використання в управлінських рішеннях.

Після того, як рейтинги сформовано, наступним кроком є їхній детальний аналіз. Цей процес дозволяє не тільки інтерпретувати результати оцінювання, а й глибше зрозуміти, як саме різні фактори взаємопов'язані. Через аналіз можна виявити, які аспекти діяльності потребують додаткових зусиль для удосконалення, які стратегії виявилися найбільш ефективними, а також як зміни в одній області можуть вплинути на результати в інших.

Рейтинги відіграють важливу роль у багатьох аспектах управління та стратегічного розвитку компаній і організацій. Вони використовуються для вирішення широкого кола управлінських завдань. Основною метою використання рейтингів є надання керівництву ефективного інструменту для прийняття обґрунтованих управлінських рішень. Ось кілька прикладів їх використання:

- У контексті стратегічного плану вони дозволяють оцінити слабкі та сильні сторони, а також можливості та загрози зовнішнього середовища, що є досить важливим для розробки ефективних стратегій у майбутньому.
- У сфері управління ризиками рейтинги служать для розпізнавання та аналізу ризиків, які можуть вплинути на бізнес. Ці рейтинги також дають можливість керівництву розробляти стратегії мінімізації ризиків і готовності до можливої кризової ситуації.
- Рейтинги також є дуже важливим інструментом внутрішнього аудиту. Вони дозволяють оцінити ефективність внутрішніх процесів і систем управління та виявити ті сфери, які потребують вдосконалення або виправлення.

Використання рейтингів дозволяє керівництву не тільки визначати пріоритети, але й формувати політики та оцінювати результативність різноманітних ініціатив і інвестицій, що є фундаментальним для забезпечення сталого розвитку і конкурентоспроможності в динамічному бізнес-середовищі.

Наприклад, у роботі [3] автор описує революційний підхід до рейтингування, який інтегрує процесне управління у формування рейтингів для інноваційних суб'єктів малого підприємства (СМП). Цей метод підкреслює важливість гнучкості у рейтингових процесах, адаптованих до швидкозмінного середовища інноваційного бізнесу. Рейтинг, за думкою автора, має служити не тільки як міра оцінки, а й як інструмент консультації, допомагаючи СМП зорієнтуватися у фінансовому світі та підвищити їхню інвестиційну привабливість.

Не тільки у бізнесі, а і в державному секторі, рейтинги виступають одним з ключових інструментів оцінки та управління діяльністю організації. Вони допомагають оптимізувати процеси та забезпечують достатній рівень прозорості та

підзвітності, що в умовах сучасної економіки є критично важливим для динамічного існування.

У бізнес-середовищі рейтинги використовуються переважно для аналізу кредитоспроможності компаній. Таким чином, вони дають можливість інвестору або кредитуру оцінити рівень ризиків, пов'язаних з вкладенням коштів у ці компанії або наданням кредитів. Крім того, рейтинги можуть застосовуватися в державному секторі для оцінки ефективності управління місцевими органами влади. Це важливо для встановлення того, наскільки ефективно використовуються управлінські ресурси при наданні послуг громаді. Таким чином, рейтинги можуть впливати на рішення щодо розподілу бюджетних коштів та ініціювання реформ, забезпечуючи підвищення якості публічних послуг.

Прозорість та підзвітність, яку забезпечують рейтинги, є невід'ємною умовою сучасної економіки. Вони дають можливість розвивати більш довірливі відносини між організаціями та стейкхолдерами, знижувати рівень корупції та підвищувати ефективність їхніх управлінських рішень. Такий підхід сприяє створенню здорового та конкурентного бізнес-середовища та підтримує сталий розвиток суспільства.

Економічні цикли надзвичайно впливають на якість рейтингів, особливо в контексті фінансових та інвестиційних рішень. Тому розуміння того, як зміни в макроекономічному середовищі впливають на рейтинги, є важливим для інвесторів, кредиторів та осіб, що приймають політичні рішення. Це можуть бути зміни в економічному зростанні, рівні інфляції, безробіття або навіть зрушення в політичній стабільності та регуляторному середовищі, які в сукупності впливають на об'єктивність, точність та надійність рейтингів.

Рейтинги часто мають антициклічний характер. Більше того, дослідження показують, що під час економічних спадів, коли ризики дефолту зростають, а фінансові умови стають складнішими, рейтингові агентства та інвестори застосовують більш обережний підхід. Це може призвести до більш консервативних, а отже, потенційно більш точних оцінок. І навпаки, під час

економічних бумів, коли умови виглядають сприятливими, існує ризик інфляції рейтингів через недостатнє сприйняття ризиків [4].

Економічна нестабільність та волатильність мають значний вплив на якість рейтингових оцінок. За таких умов рейтингові агентства можуть зіткнутися зі значними проблемами при прогнозуванні фінансових показників компаній або країн, що часто призводить до неточностей у рейтингах. Під час фінансових криз та політичних потрясінь, коли майбутній економічний розвиток важко передбачити, рейтинги можуть змінюватися дуже швидко, відображаючи нову інформацію та переоцінку ризиків. Такі зміни в регуляторному середовищі матимуть значний вплив на якість рейтингів. У банківському секторі, а особливо на фінансових ринках, будь-який новий закон або нормативний акт, такий як Базель III або Директива про Платоспроможність II, може вимагати від рейтингових агентств адаптації своїх рейтингових методологій. Ці зміни спрямовані на врахування нових умов кредитування, капіталізації та інших важливих параметрів, які потребують переоцінки в методах, що використовуються для оцінювання.

Подальші зміни в рамках економічних циклів суттєво впливають на сам характер та якість рейтингів, які стають більш чи менш консервативними по відношенню до загального економічного клімату. Це робить розуміння цієї динаміки критично важливим для рейтингових агентств, інвесторів, політиків та аналітиків, які використовують ці рейтинги як основу для своїх фінансових та стратегічних рішень. У цьому контексті врахування впливу економічного циклу на рейтинги може сприяти вдосконаленню методологій, підвищенню точності прийняття інвестиційних та фінансових рішень, а також забезпеченню економічної стійкості.

Усвідомлення впливу цих факторів на рейтинги робить оцінку ризиків та можливостей більш адекватною, дозволяє здійснювати ефективне управління ризиками та планувати майбутні інвестиції. Такий підхід допомагає не тільки зберегти фінансову стабільність, але й підвищити довіру інвесторів та довіру ринку, що є життєво важливим для глобальної економіки.

Також дуже важлива роль рейтингів це прийняття рішень у багатьох сферах, від фінансових установ до державного управління та споживчого вибору. Вони надають цінну інформацію, яка може бути використана для визначення ризиків, кредитоспроможності споживачів, інвестиційних перспектив, тощо [5].

У фінансовому секторі рейтинг, присвоєний такими агентствами, як Moody's, Standard & Poor's та Fitch, дуже важливий для оцінки ризику при кредитуванні компаній та урядів. Ці рейтинги допомагають кредиторам оцінити ймовірність своєчасного обслуговування боргів, що, в свою чергу, впливає на процентні ставки та умови кредитування.

Для інвесторів рейтинги є важливим інструментом у визначенні потенційної прибутковості та ризику, пов'язаного з інвестиційними проектами. Наприклад, облігації з найвищим рейтингом, як правило, є цінними паперами "інвестиційного класу" і несуть менший ризик, тоді як облігації з низьким рейтингом (спекулятивні) можуть пропонувати вищу дохідність, але з більшим ризиком [6].

Кредитні рейтинги також впливають на регуляторні вимоги, особливо в банківській сфері. Базельська угода передбачає, що банки повинні мати рівень капіталу відповідно до кредитного рейтингу своїх активів. Це може зменшити потребу в резервах капіталу при вищих рейтингах, або, навпаки, при нижчих рейтингах банки будуть змушені збільшувати резерви для покриття потенційних збитків.

У сфері споживчих товарів та послуг рейтинги можуть відігравати величезну роль у впливі на сприйняття бренду та поведінку споживачів. Наприклад, високі рейтинги готелю чи ресторану на TripAdvisor або Google можуть залучити більше клієнтів, тоді як низькі рейтинги можуть відштовхнути потенційних відвідувачів [7].

Управління на основі рейтингів пропонує значні можливості для підвищення ефективності бізнесу та управлінських процесів, однак також стикається з серйозними викликами, які стосуються точності, сприйняття, та етики використання рейтингів. Ось декілька ключових аспектів цих викликів та можливостей.

Виклики в рейтинговому управлінні:

- Суб'єктивність та упередженість: Рейтинги часто базуються на людських оцінках, що може призвести до суб'єктивних висновків та упереджень. Наприклад, рейтинги фільмів або ресторанів, які засновані на особистих уподобаннях рецензентів, можуть не відображати загальну якість.
- Маніпуляції та ігрові механізми: Організації можуть намагатися "грати систему" для поліпшення своїх рейтингів, адаптуючи свою діяльність під критерії оцінювання. Це може включати стратегії, які приховують реальні проблеми або недоліки.
- Залежність від рейтингів: Занадто сильна залежність від рейтингів у прийнятті рішень може призвести до недооцінки інших важливих факторів, які важко квантифікувати, наприклад, культурні чи інноваційні аспекти діяльності.

Можливості в рейтинговому управлінні:

- Підвищення прозорості та об'єктивності: Рейтинги можуть надати прозору та об'єктивну основу для оцінки продуктів, послуг, і навіть корпоративного управління, підвищуючи довіру споживачів та інвесторів.
- Підтримка стратегічного планування: Рейтингові дані можуть використовуватися для аналізу тенденцій, виявлення слабких місць та визначення пріоритетних напрямків розвитку, допомагаючи керівництву компаній формувати ефективні стратегії.
- Покращення ринкової конкуренції: Високі рейтинги можуть стимулювати компанії підвищувати якість своїх товарів або послуг, сприяючи здоровій конкуренції та інноваційному розвитку.

Цей аналіз показує, що хоча рейтингове управління має свої складнощі, воно також пропонує значні можливості для підвищення ефективності та інноваційного розвитку в бізнесі.

Отже, рейтингове управління грає велику роль у сучасній економіці, надаючи керівництву, інвесторам та споживачам об'єктивні інструменти для оцінки різноманітних аспектів діяльності. Ці оцінки сприяють прозорому прийняттю рішень, допомагають управляти ризиками, формувати стратегії та вибирати

інвестиції на основі кількісно оцінених параметрів. Впровадження рейтингового управління дозволяє компаніям підвищувати свою конкурентоспроможність, а також стимулює прозорість і об'єктивність у бізнесі та публічному управлінні.

1.2 Особливості кіноіндустрії та необхідність рейтингування фільмів

Кіноіндустрія, яка є однією з найбільш значущих та впливових сфер світової економіки та культури, охоплює весь спектр від виробництва фільмів до їх розповсюдження у різних форматах і на різних платформах, таких як кінотеатри, телебачення, та цифрові стрімінгові сервіси [8].

Ця галузь має глибокий культурний вплив, оскільки фільми формують суспільні норми та вартості через візуальні наративи, які вони пропонують. Фільми не лише розважають; вони також служать дзеркалом чи критичним аналізом суспільних явищ, відображаючи і впливаючи на суспільні настрої та ідеології. Вони піднімають важливі соціальні питання, стимулюють обговорення та спонукають до суспільних змін, формуючи громадську думку про різні події та особистості.

З економічної точки зору, кіноіндустрія є джерелом численних робочих місць — від акторів та режисерів до кінотеатрів та дистриб'юторів. Вона вносить значний вклад у податкові надходження багатьох країн та сприяє розвитку туризму, оскільки популярні локації у фільмах часто стають туристичними магнітами. Крім того, пошук нових способів зйомки та показу фільмів стимулює технологічний розвиток, особливо у галузях звукових та візуальних технологій, відкриваючи нові можливості для інновацій та креативності.

Далі оглянемо ключові аспекти кіноіндустрії. Виробництво фільму — складний процес, який вимагає ретельного планування та бездоганної координації. Все починається зі сценарію, що є основою кожного кінопроекту. Цю творчу ідею розробляють сценаристи, після чого вона відтворюється на екрані у візуальній формі. Після створення сценарію продюсери та режисери переходять до

планування бюджету, акторського складу, реквізиту та локацій. Кожен окремий аспект є критичним і важливим для управління ресурсами та питанням часу. Кожен кадр фільму створюється за участю операторів, акторів та звукорежисерів. Монтаж та постпродакш – останні стадії процесу, коли фільм приймає остаточний вигляд. Тим часом відбувається монтаж відзнятого матеріалу зі спецефектами, а також монтаж звуку та фінальне зведення. Кожен процес потребує керівництва і уваги до деталей, оскільки якість кінцевого продукту визначить його успіх серед глядачів та критиків. На кожному етапі кіновиробництва суми інвестицій є дійсно значними, а сам процес, у свою чергу, також вимагає чіткої взаємодії з усіма учасниками. Це не лише велика творча робота, а й, з іншого боку, досить складна логістична операція, успіх якої може мати досить значний вплив на фінансове та культурне життя кіноіндустрії [9].

Рейтингування фільмів виконує критичну функцію у кіноіндустрії, захищаючи глядачів, служачи свого роду нормативами та впливаючи на комерційний успіх фільмів.

Кінореєтинги відіграють вирішальну роль у кіноіндустрії, виконуючи кілька важливих функцій. Перш за все, вони захищають глядачів, надаючи необхідну інформацію про зміст фільму. Це дозволяє людям робити усвідомлений вибір щодо того, що дивитися їм або їхнім дітям. Це особливо важливо, коли йдеться про контент, який може містити насильство, сексуальні сцени, вживання наркотиків чи ненормативну лексику [10].

У Сполучених Штатах діє рейтингова система для фільмів, якою керує Американська асоціація кінематографістів (МРАА), що включає в себе такі категорії:

- G (Загальна аудиторія): Ці фільми підходять для всіх вікових груп. Вони не містять матеріалів, які можуть вважатися неприйнятними для дітей. Фільми з рейтингом G, як правило, не містять насильства, сексуального змісту, вживання наркотиків та грубої лексики.

- PG (Рекомендується присутність батьків): Ці фільми можуть містити сцени, які не підходять для перегляду дітьми без нагляду дорослих. Вони можуть

включати легке насильство, провокаційну лексику, короткі зображення наркотиків або легкі сексуальні сцени. Рейтинг PG означає, що батьки, можливо, захочуть подивитися фільм першими або у супроводі своїх молодших дітей.

- PG-13 (Наполегливе застереження батькам): Фільми з таким рейтингом можуть містити сцени насильства, крові, сексуального змісту або вживання наркотиків, які можуть бути неприйнятними для дітей до 13 років. Рейтинг PG-13 радить батькам бути особливо обережними і подумати, чи варто дивитися фільм дітям до 13 років.

- R (Обмежений): Дітям до 17 років рекомендується дивитися фільми з цим рейтингом тільки в компанії дорослих. Ці фільми можуть містити сильні сцени насильства, ненормативну лексику, сексуальний зміст або інтенсивне вживання наркотиків. Рейтинг R вказує на те, що зміст фільму має характер, який більшість батьків вважатимуть занадто дорослим для дітей до 17 років.

- NC-17 (Тільки для дорослих): Ці фільми містять дуже відвертий зміст і, як правило, не підходять для глядачів молодше 17 років. Фільми з рейтингом NC-17 можуть містити відвертий сексуальний зміст, сцени насильства або графічні зображення вживання наркотиків. Цей рейтинг вказує на те, що фільм не можна показувати особам молодше 17 років [11].

Окрім захисту, рейтинги виступають своєрідним регуляторним стандартом, встановлюючи орієнтири для контенту, який виробляється та демонструється. Вони роблять кіноіндустрію підзвітною суспільству, гарантуючи, що фільми відповідають певним культурним і соціальним очікуванням.

Але рейтинги - це не лише захист чи стандартизація. Вони також суттєво впливають на комерційний успіх фільму. Високі рейтинги можуть залучити більшу аудиторію до кінотеатрів, збільшуючи касові збори, тоді як низькі рейтинги можуть відлякати потенційних глядачів. Маркетологи також використовують рейтинги, щоб орієнтувати фільми на певну аудиторію, адаптуючи свої промо-стратегії для максимального охоплення та впливу.

Таким чином, система кінорейтингів слугує важливим інструментом в управлінні споживчим досвідом, стандартизації контенту та забезпеченні

комерційного успіху кінопродукції. Це не просто цифра, це орієнтир, який допомагає орієнтуватися у великому та різноманітному ландшафті кіно.

Рейтинги фільмів допомагають не лише зорієнтуватися глядачеві у змісті, але й стали одним із ключових елементів маркетингової стратегії кіноіндустрії. Завдяки цьому збільшується не лише відбір контенту, але й формування очікувань глядачів, особливо щодо фільмів з рейтингом R, що часто приваблює молодь, яка шукає гострих відчуттів. Цей рейтинг часто приваблює молодь, яка шукає емоційного збудження або бажає відчувати себе дорослішою, "спробувавши щось заборонене" [12].

Концепція "забороненого плоду" впливає на молоду аудиторію. Цей ефект базується на психологічному інстинкті, за яким заборонені речі сприймаються як більш привабливі. Згідно з цією концепцією заборонені речі завжди спонукають людей йти за ними. Фільми з рейтингом R часто містять сцени, які виходять за межі уяви кіноглядачів. Такі елементи можуть сприйматися як особливо цікаві для молоді. Використання рейтингу R у маркетингових стратегіях дозволяє позиціонувати фільми як більш серйозні чи глибокі. Маркетингові кампанії можуть підкреслити унікальність фільму таким чином, що він дійсно може викликати сильні емоції або подарувати щось незабутнє. Це заохочує інтерес не лише молоді, а й більш досвідченої аудиторії, яка цінує складність і багатство кінематографічного мистецтва.

Фільми з таким рейтингом часто демонструють високі касові збори, адже "заборонений" контент приваблює широку аудиторію, яка готова платити за досвід, що виходить за традиційні рамки. Не менш важливу роль у цьому відіграє вірусний маркетинг: обговорення в соціальних мережах, рецензії та словесний маркетинг значно збільшують популярність фільму.

Рейтинги фільмів мають вирішальне значення для комерційного успіху кінопроектів, оскільки вони, як правило, мають відносно значний вплив на касові збори. З усіх класів фільми з рейтингом PG-13 часто мають найвищий рейтинг, що виправдано їхньою здатністю залучати найширшу глядацьку аудиторію. По суті, фільми PG-13 досить добре балансують між сімейною та дорослою тематикою, що

дозволяє їм охоплювати і залучати як підлітків, так і дорослих. Вони можуть містити відвертий контент про насильство, сексуальність, зловживання наркотиками чи алкоголем, але не настільки, щоб це відштовхувало від них сімейну аудиторію. Це дозволяє потенційно збільшити кількість глядачів, серед яких може бути і молодша аудиторія [13].

PG-13 дозволяє маркетинговим командам використовувати ширший спектр стратегій для просування фільмів. Такі фільми можна агресивно рекламувати не лише на телебаченні та в Інтернеті, а й через партнерство з такими брендами. Крім того, залучення кумирів поп-культури та інфлюенсерів до рекламних кампаній також посилює вплив і привабливість фільму навіть серед такої вікової аудиторії.

Багато фільмів з рейтингом PG-13 часто містять елементи, які приваблюють різні вікові групи, щоб забезпечити вражаючі касові збори. Підлітки та молодь сприймають такі фільми серйозно, тоді як дорослі оцінюють здатність фільму залишатися достатньо стриманим для перегляду разом з молодшими членами сім'ї. Це означає, що фільми PG-13 надихають більше відвідувачів кінотеатрів, оскільки вони містять зміст, прийнятний для всіх видів дружніх зустрічей - від простих посиденьок до сімейних прогулянок. Це додає до загальної кількості переглядів і повторних відвідувань, оскільки подібні фільми є саме такими.

У сукупності ці фактори роблять фільми з рейтингом PG-13 не лише популярними, але й надзвичайно прибутковими з комерційної точки зору, що призводить до великих касових зборів як на національному, так і на міжнародному рівнях.

Рейтинги фільмів суттєво впливають на сприйняття та вибір споживача, виконуючи не лише інформаційну функцію, але й виступаючи маркером якості, який може окреслити успіх чи провал фільму на ринку. Таким чином, вплив на сприйняття якості, а також на регулювання та стандартизацію в кіноіндустрії є багатовимірним і глибоко вплетеним у механіку кіноринку.

Високі рейтинги на різних платформах, таких як IMDb, Rotten Tomatoes чи Metacritic, свідчать про високу якість фільмів. Це може привабити більше глядачів до кінотеатру і призвести до вибору певного фільму для перегляду. Позитивні

очікування виникають через високі рейтинги, а низькі можуть призвести до відмови від ідеї переглянути фільм, навіть якщо він має інші привабливі аспекти, такі як улюблений режисер чи актори. Це може бути пов'язано з теорією соціального доказу, коли люди, ймовірно, більше залежать від дій і навіть думок інших людей [14].

Серед іншого, споживачі використовують рейтинги як ярлик, щоб швидко визначити, що подивитися. Обираючи фільми з високими рейтингами, глядачі мають позитивні очікування та позитивний досвід. Сильніші очікування від високих рейтингів також породжують позитивні очікування. Цей ефект відомий як самовідтворне пророцтво, що самоздійснюється, оскільки дозволяє фільмам з високим рейтингом привернути більше уваги та отримати позитивну оцінку в майбутньому.

Рейтингові системи є основою для забезпечення регулювання та стандартизації кіноіндустрії, щоб забезпечити послідовність і передбачуваність класифікації фільмів. Це важливо для продюсерів, дистриб'юторів та глядачів по всьому світу. Уніфікація рейтингів, безумовно, допомогла б, як це робить система МРАА у США чи BBFC у Великій Британії, щоб глядачі знали, з чим мають справу, незалежно від того, де вони дивляться фільм. Це сприяє міжнародному культурному обміну та захисту молоді, забезпечуючи належне регулювання контенту, який може бути шкідливим або неприйнятним для певних вікових груп.

Крім того, рейтинги впливають на презентацію фільмів у кінотеатрах і на телебаченні, а також на спосіб їхньої реклами. У країнах із найсуворішим контролем і найсуворішими органами у сфері кінематографії, таких як Японія чи Південна Корея, рейтинги визначають, коли фільм може бути придатним для перегляду. Як наслідок, кіноіндустрія покладається на регуляторний контроль за часовими інтервалами або навіть регіонами, в яких фільм має бути доступним, що сприяє перевірці регуляторних і культурних стандартів. У зв'язку з цим рейтинги фільмів не лише формують дохідну частину бюджету завдяки сприйняттю, яке формується у споживача, але й сприяють структуруванню та регулюванню кіносектору [15].

Отже, можемо сказати що рейтингування фільмів є невід'ємним інструментом в кіноіндустрії з кількох причин. По-перше, воно захищає молодь від контенту, який може бути для них недоречним або шкідливим, обмежуючи доступ до матеріалів з високим рівнем насильства або сексуального змісту. По-друге, рейтинги впливають на комерційний успіх фільмів, допомагаючи виробникам та дистриб'юторам визначити потенційну аудиторію і оптимізувати маркетингові стратегії. По-третє, вони допомагають споживачам ухвалювати обґрунтовані рішення про те, що дивитися, надаючи чітку інформацію про вміст фільму. Нарешті, рейтинги виконують важливу регулятивну функцію, стандартизуючи критерії оцінки фільмів у глобальному масштабі, що сприяє уніфікації індустріальних стандартів і сприяє культурному обміну між країнами.

1.3 Особливості IMDb та чинники формування його рейтингу

Кінорейтинги відіграють важливу роль у кіноіндустрії, оскільки впливають на вибір глядачів, інвестиційні рішення студій та навіть визначають загальну оцінку фільму. Серед багатьох доступних платформ для оцінювання фільмів IMDb (Internet Movie Database) виділяється як одна з найпотужніших.

Заснований у 1990 році, IMDb вперше побачив світ завдяки групі шанувальників кіно в дискусійній групі новин Usenet під назвою rec.arts.movies. Вони мали на меті структурувати існуючий FAQ, який, по суті, містив кілька тисяч статей про близько десяти тисяч фільмів та серіалів. Один із засновників, який згодом став виконавчим директором IMDb, Кол Нідем, написав низку UNIX-скриптів, які згодом стали корисними як для структурування, так і для редагування бази даних. Від самого початку IMDb розробив свою систему на основі відкритого програмного забезпечення, дозволивши широкому колу розробників стати частиною проекту [16].

Коли база даних вперше з'явилася в Інтернеті в 1993 році, вона дебютувала на сервері валлійського університету Кардіффа і одразу ж була помічена. Тоді середня кількість запитів на місяць становила близько 65 мільйонів.

Вже в 1998 році ситуація кардинально змінилася: Джеф Безос із Amazon зробив пропозицію, повністю відповідну духу IMDb. Безос пропонував фінансову підтримку, одночасно залишаючи доступ до ресурсів бази безкоштовним для всіх користувачів. Його єдиною вимогою було право Amazon.com на продаж кінопродукції через сайт IMDb.com, а також право використовувати цей сайт як рекламну платформу для своїх товарів. Така пропозиція підійшла обом сторонам: IMDb змогла одержати необхідні ресурси для розширення та підтримки своєї ростучої бази даних, тоді як Amazon.com отримала можливість використовувати популярність IMDb для збільшення продажів [17].

Завдяки партнерству з Amazon технічні можливості IMDb зросли в рази, що призвело до фінансової стабільності, необхідної для підтримки безперервного розвитку серверних потужностей. Це також дозволило IMDb найняти багато співробітників, значна частина з яких були ентузіастами сервісу IMDb. У свою чергу, відома та популярна платформа дозволила Amazon просувати свої товари, а отже, збільшити її присутність на ринку інтернет-торгівлі.

Обидві компанії виявили, що співпраця є вигідним процесом, оскільки вона зберігає ключові цінності IMDb – безкоштовний доступ для аудиторії до величезної бази даних про кіно та акторів, одночасно забезпечуючи комерційний успіх завдяки використанню продукції Amazon. Чим більший функціонал отримував сайт, тим різноманітнішими ставали можливості його користувачів: від рейтингування фільмів та написання рецензій до персональних списків фільмів та участі у форумах. Рейтинги IMDb стали важливою річчю для кіноіндустрії у виборі фільмів для перегляду публікою та оцінці фільмів критиками.

Таким чином, IMDb перетворився з простого проекту групи ентузіастів кіно на глобальний ресурс, який продовжує надавати мільйонам людей по всьому світу актуальну інформацію і водночас виступає мостом між кіноіндустрією та громадськістю.

В 2002 році IMDb запровадила підписковий сервіс під назвою IMDbPro, який став значним кроком у розвитку компанії та її послуг. IMDbPro був розроблений спеціально для професіоналів кіноіндустрії, включаючи акторів, режисерів, продюсерів та інших фахівців, які потребували більш глибокого доступу до індустріальної інформації, що надавалась IMDb. Цей сервіс був створений для того, щоб забезпечити фахівцям індустрії необхідні інструменти для роботи, а також дозволити їм краще орієнтуватися в галузевій інформації, яка була важлива для їх професійної діяльності.

IMDbPro надає доступ до розширених функцій пошуку та детальної контактної інформації про індустрію, події, кінопроекти та персональні профілі на IMDb. Основні функції IMDbPro включають:

- Деталізовані профілі акторів, режисерів та сценаристів, а також контактну інформацію їхніх агентів та менеджерів.
- Надає інформацію про касові збори, що дозволяє користувачам аналізувати комерційний успіх фільмів.
- Інформує професіоналів про останні події та новини кіноіндустрії.
- Пропонує можливість відстежувати етапи розвитку різних кінопроектів, що має вирішальне значення для тих, хто шукає можливості працевлаштування або співпраці в індустрії [18].

Запуск IMDbPro став знаковою подією для кіноіндустрії, надавши професіоналам потужний інструмент для більш ефективної роботи. Підписка дозволяє користувачам глибше зануритися в аналітику та управління кар'єрою, які раніше були доступні лише через внутрішні канали індустрії. Це перетворило IMDb не лише на платформу для перегляду фільмів та оцінки їхніх рейтингів, а й на важливий ресурс для професіоналів кіноіндустрії. Інформація, яка раніше могла бути прихованою або важкодоступною для новачків кіноіндустрії, перетворилася на щось дуже доступне для широкого кола професіоналів.

З появою IMDbPro також змінився спосіб взаємодії в мережі в кіноіндустрії. Платформа дозволяє користувачам встановлювати контакти з потенційними партнерами або роботодавцями, дізнаватися більше про колег по галузі, а також

виявляти нові таланти. Можливість для нетворкінгу стала невід'ємною частиною професійної діяльності, сприяючи створенню нових професійних зв'язків та підтриманню існуючих. Сервіс надає акторам можливість відстежувати нові проекти та зв'язуватися з кастинг-директорами, щоб підвищити свої шанси на отримання ролі. Інструменти аналітики IMDbPro дозволяють вимірювати популярність акторів на основі переглядів їхніх профілів та руху рейтингу "STARMeter", що може вплинути на їхню бажаність для кастингу в нових проектах.

Відгуки про IMDbPro у спільноті загалом позитивні, оскільки сервіс постійно адаптується до потреб користувачів і реагує на зміни в індустрії. Завдяки регулярному оновленню функціоналу, покращенню інтерфейсу та впровадженню нових можливостей IMDbPro залишається цінним ресурсом для професіоналів кіноіндустрії.

У майбутньому очікується подальша інтеграція IMDbPro з новими технологічними рішеннями та платформами, зокрема зі штучним інтелектом та машинним навчанням. Ця інтеграція забезпечить глибший аналіз даних та персоналізовані рекомендації для користувачів. Такі досягнення виводять IMDbPro на передові позиції в кіноіндустрії, сприяючи інноваціям та кар'єрному зростанню в цій динамічній та захоплюючій галузі.

На березень 2024 року в базі IMDb було приблизно 19 мільйонів назв, у тому числі фільми, телесеріали, музичні відео, короткометражні фільми, а також інформація про 200 мільйонів осіб, що об'єднують фахівців кіноіндустрії, включаючи акторів, режисерів, сценаристів та інших. Також станом на 2022 рік сервіс IMDb налічував 83 мільйони зареєстрованих користувачів. Але велика кількість користувачів продовжує користуватись їм без реєстрації, що підтверджується великою кількістю активних незареєстрованих користувачів та переглядами сайту. За березень 2024 року нараховується понад 520 мільйонів[19, 20].

Рейтинг IMDb формується на основі голосів зареєстрованих користувачів. Кожен фільм можна оцінити за шкалою від 1 до 10, а остаточний рейтинг виводиться як середнє арифметичне всіх зареєстрованих голосів. Для розрахунку

рейтингу, IMDb використовує складну формулу під назвою "зважене середнє", яка допомагає запобігти спотворенням і маніпуляціям з рейтингом. Вона використовується для обчислення кінцевого рейтингу кожного фільму на основі оцінок, виставлених користувачами [12, 21].

Основна формула, яку IMDb використовує для визначення зваженого середнього рейтингу фільму (Weighted Rating, WR), може бути представлена наступним чином:

$$WR = \left(\frac{v}{v + m} \cdot R \right) + \left(\frac{m}{v + m} \cdot C \right),$$

де R - середній рейтинг фільму, тобто середньою оцінкою, яку фільм отримав від користувачів;

v - представляє кількість голосів, які фільм отримав від користувачів;

m - порогове значення, мінімальна кількість голосів, необхідна для того, щоб фільм потрапив до рейтингу. IMDb встановлює це число, щоб забезпечити достатню вагу голосів у рейтингу (зазвичай 25,000 голосів для рейтингу Top 250);

C - середній рейтинг усіх фільмів на платформі.

Використання зваженого середнього рейтингу є ключовою стратегією IMDb для забезпечення точності та реалістичності у процесі оцінювання фільмів. Цей метод обчислення рейтингу враховує не лише середні оцінки, а й кількість отриманих голосів, що суттєво впливає на остаточний рейтинг фільму на платформі.

Хоча зважене середнє є ефективним методом підвищення точності рейтингів, воно також має свої обмеження та проблеми. Наприклад, нові фільми, які ще не набрали достатньої кількості голосів, можуть показувати хибний рейтинг, доки не досягнуть певного порогу m . Це може спотворити реальну картину їхньої популярності або якості незабаром після виходу в прокат.

Для подальшого вдосконалення рейтингової системи IMDb міг би розглянути варіанти більш динамічного коригування порогового значення m , залежно від категорії фільму або швидкості збору голосів. Крім того, можливе використання більш досконалих аналітичних інструментів для виявлення незвичних моделей голосування, які можуть свідчити про спроби маніпулювання рейтингами. Ці та

інші вдосконалення можуть допомогти зміцнити довіру користувачів до рейтингів IMDb, забезпечивши їхню актуальність та відповідність сучасним потребам користувачів та кіноіндустрії.

Окрім формули по розрахунку рейтингу є інші чинники, які впливають на рейтинг фільму. Фактори, що впливають на рейтинги фільмів на IMDb, демонструють складність та динамічність взаємодії користувачів з контентом. Ось кілька основних з них:

- **Кількість голосів:** Кількість голосів, яку отримує фільм, безпосередньо впливає на стабільність його рейтингу на IMDb. Фільми з великою кількістю голосів, як правило, мають стабільніший рейтинг, оскільки окремі екстремальні оцінки мають менший вплив на загальний рейтинг. Це також свідчить про широкий глядацький інтерес, який може бути зумовлений різними факторами, зокрема успіхом фільму на фестивалях, його промо-кампанією або участю популярних акторів. Крім того, велика кількість голосів також може свідчити про значну глядацьку аудиторію, що є показником популярності фільму [22].

- **Відгуки та рецензії:** Рецензії та критика відіграють вирішальну роль у формуванні громадської думки про фільм. Позитивні рецензії можуть значно підвищити інтерес до фільму, тоді як негативні можуть відлякати потенційних глядачів. Рецензії від впливових критиків або популярних рецензійних платформ можуть мати ще більший вплив, оскільки вони часто формують очікування глядачів ще до того, як вони побачать фільм. Рецензії також стимулюють обговорення та подальший обмін думками в Інтернеті, що сприяє підвищенню інтересу до фільму та його аналізу [23].

- **Соціальні медіа та інтернет-активність:** Обговорення фільмів у соціальних мережах та на інших інтернет-платформах суттєво впливає на рейтинги IMDb. Онлайн-активність може швидко розширити аудиторію фільму, особливо серед молодих глядачів, які активно користуються соціальними мережами. Вірусні відео, меми, хештеги та інші форми онлайн-взаємодії значно підвищують впізнаваність фільму і можуть заохотити глядачів подивитися його та залишити свої оцінки на IMDb. Наприклад, культурні явища, які набувають популярності в

соціальних мережах, часто зазнають значного зростання рейтингів завдяки посиленому обговоренню та обміну досвідом в Інтернеті. Цей процес демонструє, як соціальні мережі можуть виступати каталізатором інтересу до фільму, впливаючи на його загальне сприйняття та рейтинги [24].

- Час випуску та актуальність: Період виходу та актуальність фільму також відіграють значну роль у формуванні його рейтингу на IMDb. Нові фільми часто отримують більше уваги та вищі рейтинги одразу після виходу через початковий ажіотаж та маркетингові кампанії. Однак з часом ці рейтинги можуть змінюватися, оскільки все більше людей дивляться фільм і залишають свої відгуки, що відображає довгострокове сприйняття фільму аудиторією. Це підкреслює динамічну природу рейтингів IMDb, де початкові оцінки не завжди свідчать про довготривалу цінність та якість фільму [25].

- Культурні та регіональні особливості: Культурні та регіональні особливості можуть суттєво впливати на рейтинги фільмів на IMDb. Фільми, що висвітлюють культурно-специфічні теми або орієнтовані на певну географічну аудиторію, можуть мати різні рейтинги в різних регіонах. Наприклад, міжнародні фільми часто отримують різні оцінки від аудиторії, що належить до різних культур, залежно від того, наскільки добре теми та персонажі резонують з місцевими цінностями та переконаннями. Цей аспект особливо важливий для таких глобальних платформ, як IMDb, оскільки він відображає розмаїття глядацьких уподобань і культурних перспектив [26].

Отже, можна сказати що IMDb відіграє колосальну роль у кіноіндустрії, впливаючи на сприйняття та популярність фільмів. Ці рейтинги формуються на основі середньозваженого алгоритму, який враховує кількість голосів, середній рейтинг фільмів та активність глядацької аудиторії. Значний вплив також мають рецензії, соціальні мережі, час виходу фільму на екрани та культурні особливості регіонів. Розуміння цих факторів дозволяє краще аналізувати тенденції та реакцію аудиторії, розширюючи стратегічні можливості в маркетингу та дистрибуції фільмів.

РОЗДІЛ 2

МОДЕЛЮВАННЯ РЕЙТИНГУ ФІЛЬМІВ НА ПРИКЛАДІ РЕЙТИНГУ IMDB

2.1 Моделювання рейтингу на підґрунті кластерного аналізу

Аналіз буде проводитися за датасетом IMDb. Структуру датасету, використаного у проведенні дослідження, наведено у Додатку А. Датасет містить 6159 спостережень, з яких більшість – фільми. А саме 4446 фільмів та 1713 серіалів. Є такі стовбці (фактори) як: Name (Назва фільму/серіалу), Date (Дата випуску), Rate (Рейтинг), Votes (Кількість голосів), Genre (Жанри), Duration (Тривалість), Type (наприклад, серіал чи кінофільм), Certificate (Сертифікат вікових обмежень), Episodes (Кількість епізодів, у випадку серіалів), Nudity (Рівень наявності оголеності), Violence (Рівень наявності насильства), Profanity (Рівень наявності нецензурної лексики), Alcohol (Рівень наявності вживання алкоголю) та Frightening (Рівень наявності лякаючих сцен). З них числові дані: Date (Дата випуску), Rate (Рейтинг), Votes (Кількість голосів), Duration (Тривалість) та Episodes (Кількість епізодів, у випадку серіалів).

Кластерний аналіз є важливим інструментом статистичного аналізу даних, що дозволяє диференціювати групи схожих об'єктів. Цей метод використовується для групування об'єктів у кластери на основі внутрішньої схожості, при цьому кожна пара об'єктів з різних кластерів повинна суттєво відрізнятися. Сфери застосування кластерного аналізу варіюються від маркетингу – визначення сегментів споживачів на основі купівельного досвіду, до біології – класифікації еволюційного потенціалу виживання за допомогою генетичних зразків. У соціальних науках кластерний аналіз виявляє приховані групи в масивах даних і надає дослідникам засоби для ідентифікації та аналізу груп людей з обмеженими соціальними, економічними або психологічними критеріями. У медицині цей метод

використовується для виявлення пацієнтів з однаковими медичними профілями для створення більш скоординованих стратегій лікування та профілактики захворювань. Кластерний аналіз виявляє закономірності та тенденції у великих масивах даних, що робить його універсальним інструментом для науковців і практиків [27].

Ієрархічний кластерний аналіз є одним з найпопулярніших методів кластеризації, який використовується для систематичного групування об'єктів на основі їхньої схожості. Цей метод може застосовуватися у двох основних формах: агломеративній (об'єднуючій) та дивізивній (поділяючій).

Агломеративний підхід є більш поширеним і починається з того, що кожен об'єкт у наборі даних розглядається як окремий кластер. Потім на кожному етапі два найближчі кластери об'єднуються, поки всі об'єкти не будуть об'єднані в один кластер. Для вимірювання відстані між кластерами можна використовувати різні метрики, такі як найближчий сусід, найвіддаленіший сусід, середня відстань або метод Уорда, який мінімізує збільшення суми квадратів відстаней всередині кластерів [28].

Дивізивний метод діє у зворотному порядку: спочатку всі об'єкти розглядаються як єдиний кластер, який потім поступово розбивається на менші кластери. Цей процес триває доти, доки кожен об'єкт не стане окремим кластером. Цей метод є менш поширеним через його більшу обчислювальну складність порівняно з агломеративним методом [29].

Для візуалізації ієрархічних кластерів часто використовують дендрограму - деревоподібну діаграму, яка показує порядок об'єднання кластерів і відстані між ними. Дендрограми дозволяють легко визначити, які групи є найбільш схожими, і можуть допомогти визначити оптимальну кількість кластерів шляхом зрізання дерева на певному рівні.

Ієрархічний кластерний аналіз використовується в різних галузях, включаючи генетичні дослідження для аналізу схожості ДНК, соціальні науки для групування схожих демографічних груп і маркетинг для сегментації споживачів. Ці

застосування демонструють універсальність методу та його здатність виявляти приховані закономірності в складних наборах даних.

Також є метод К-середніх (K-means), він є одним із найпопулярніших алгоритмів кластеризації, який широко використовується у статистичному аналізі даних для розділення набору об'єктів на К передбачуваних кластерів. Основна мета алгоритму — мінімізувати суму квадратів відстаней між точками і центром кластера, до якого ці точки належать, тобто формула може виглядати так:

$$\sum_{i=1}^N d(x_i, m_j(x_i))^2,$$

де $\sum_{i=1}^N$ — сума для всіх об'єктів даних від 1 до N;

d - метрика відстані;

x_i - i -й об'єкт даних;

$m_j(x_i)$ - центр кластера, якому на j -ій ітерації приписаний елемент x_i .

Метод К-середніх є часто використовуваним у багатьох областях: клієнтська сегментація, групування текстових документів у тематичні категорії, зображення та ін. Популярність цього алгоритму визначається простістю і ефективністю, оскільки правильне налаштування моделі та якісні джерела даних можуть гарантувати хороші результати [30].

Отже, переходимо до кластерного аналізу рейтингу фільмів IMDb. Роботу я буду виконувати в середовищі R Studio, за допомогою мови програмування R.

Першочергово завантажуюмо датасет, обираємо числові дані, які будемо використовувати в аналізах та очищуємо пусті значення за допомогою функції `na.omit()`. Для подальших аналізів я вибрав такі числові дані: Рейтинг фільму, Кількість голосів та Тривалість фільму. Далі стандартизуємо дані та переходимо до ієрархічного кластерного аналізу.

Дендрограма на Рис 2.1 показує, що дані можна розумно розділити на чотири основні кластери. На це вказують великі «стовбури» на дендрограмі, які поділяють дані на відносно великі, чіткі групи. Кожен кластер представляє групу фільмів, які схожі між собою за кількома критеріями, визначеними під час кластерного аналізу. Відстань на дендрограмі між різними кластерами вказує на

ступінь відмінності між групами. Чим вище «стовбур», що з'єднує кластери, тим більша різниця між ними. Великі відстані між основними кластерами свідчать про значні відмінності в характеристиках фільмів всередині цих кластерів.

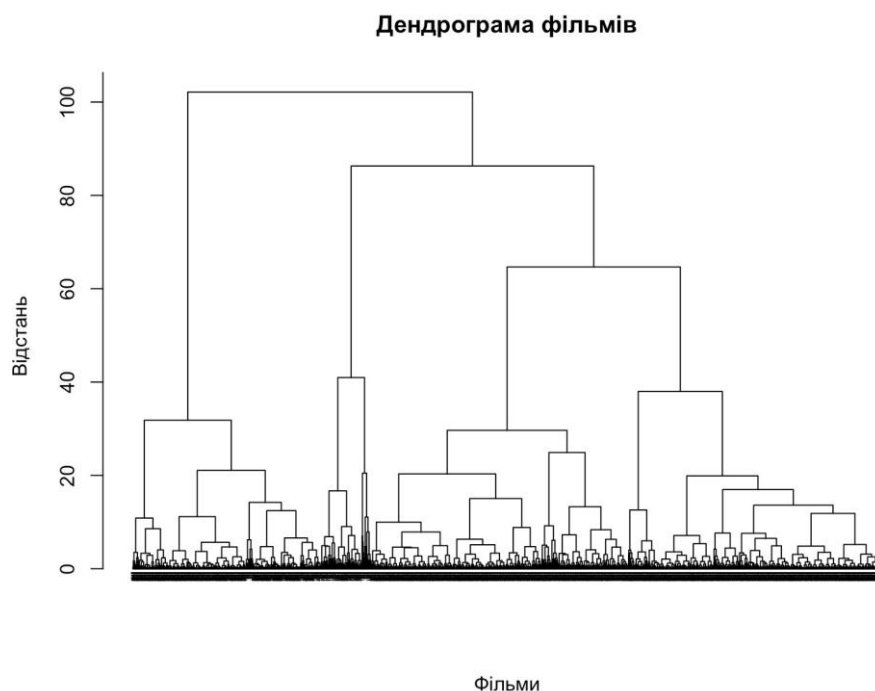


Рисунок 2.1 – Дендрограма фільмів

Джерело: складено автором.

Також далі визначимо оптимальну кількість кластерів за допомогою методу ліктя. Отримаємо такий графік (Рис 2.2).

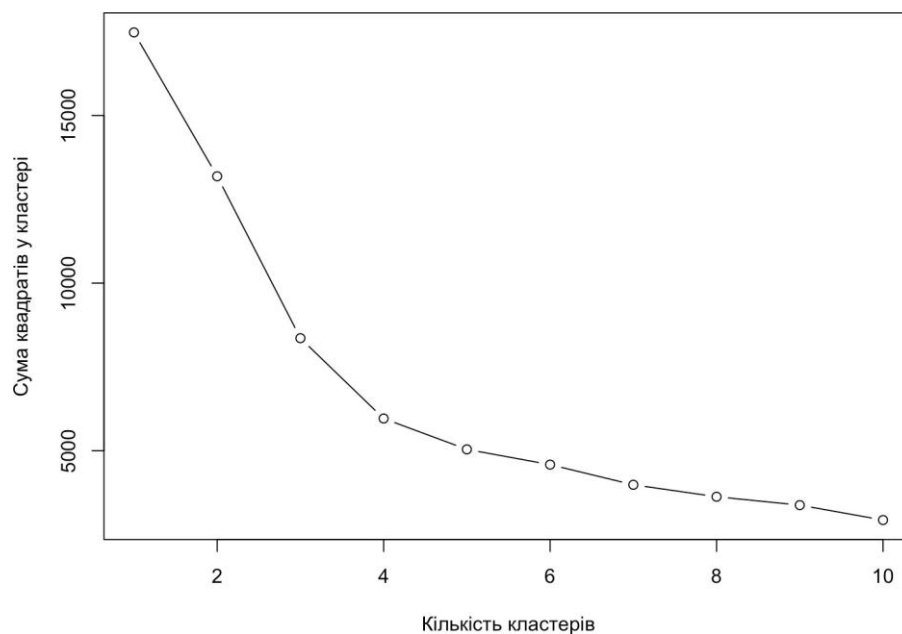


Рисунок 2.2 – Метод ліктя

Джерело: складено автором.

Дивлячись на візуалізацію методу ліктя, можна легко визначити, що внутрішньогрупова сума квадратів має тенденцію до зменшення з дуже низькою швидкістю після чотирьох кластерів. Це означає, що для нашого аналізу оптимальною кількістю кластерів має бути чотири, що ще раз підтверджує дендрограма.

Далі проводимо кластеризацію за допомогою методу k-середніх та задаємо кількість кластерів – 4. Встановлюємо початковий стан за допомогою функції `set.seed()`, щоб результати були відтворюваними. Додаємо кластерні мітки, отримані з результатів k-середніх до датасету.

Об'єднуємо дані для кожного створеного кластера та обчислюємо середнє. Після отримання даних, візуалізуємо їх. Отримуємо такий графік (Рис 2.3):

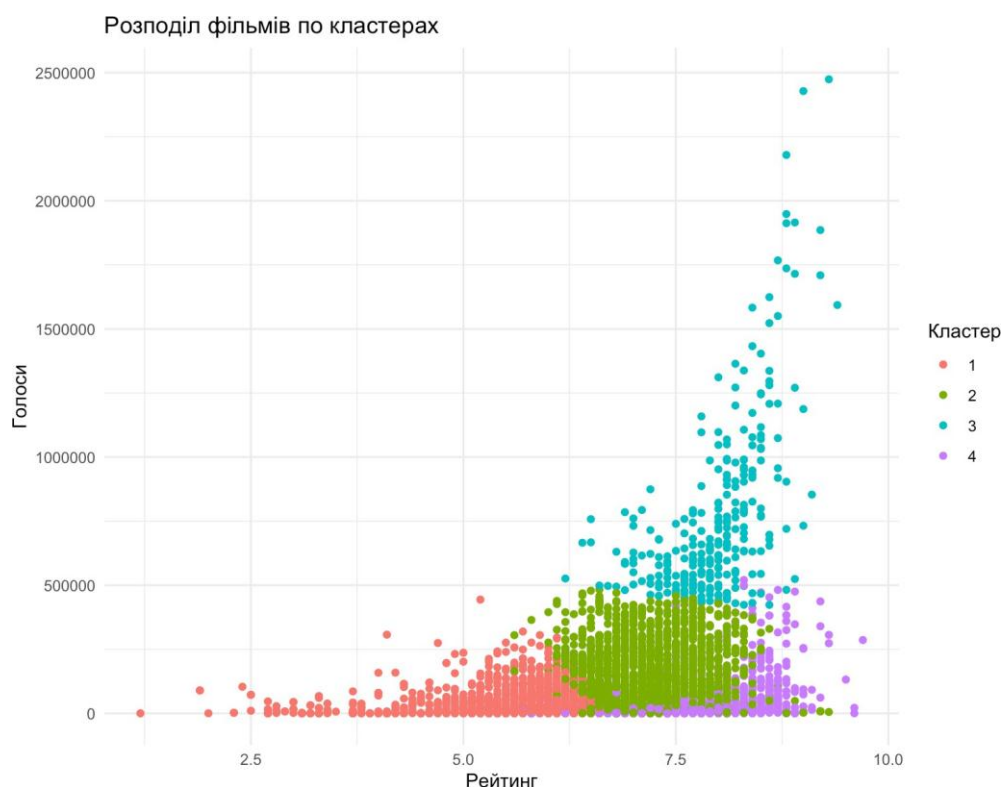


Рисунок. 2.3 - Розподіл фільмів по кластерах

Джерело: розроблено автором.

Згідно результатів, які бачимо на Рис 2.3 можемо так охарактеризувати кластери:

- Кластер 1: Включає фільми з відносно низькими рейтингами та помірною кількістю голосів, що може означати фільми меншої популярності або відносно низькі рейтинги, але середньої тривалості.

- Кластер 2: Фільми цього кластеру в середньому мають вищі рейтинги та більше голосів, що свідчить про їхню популярність та більшу тривалість. Ймовірно, вони мають більшу вагу як фільми-номінанти.

- Кластер 3: До цього кластеру належать фільми з найбільшими рейтингами, великою кількістю голосів і найбільшою середньою тривалістю показу, що належать до кластеру високореєтингових блокбастерів.

- Кластер 4: Рейтинг фільмів цього кластеру відносно високий, але тривалість і кількість голосів значно нижчі, ніж у фільмів з іншого кластеру - ймовірно, це пов'язано з коротшим форматом, наприклад, короткометражною.

Перейдемо до моделювання рейтингу фільмів на підґрунті кластерного аналізу. Спочатку розділимо дані на навчальну та тестову вибірку. Навчальна вибірка містить 4663 спостережень, що є приблизно 80% від загальної кількості записів у датасеті. Містить такі дані, як: Rate (Рейтинг), Votes (Кількість голосів), Duration (Тривалість) та Cluster (Кластер). Використовується для тренування моделі. А саме для того щоб модель "навчилася" знаходити паттерни та взаємозв'язки між змінними. Змінна Rate використовується для того, щоб забезпечити рівномірний розподіл рейтингів у навчальній вибірці.

Тестова вибірка містить приблизно 20% від загальної кількості спостережень, а саме 1164 спостереження. Використовується для оцінки ефективності та точності моделі. Подібно до навчальної вибірки, тестова вибірка містить всі ті ж дані, але у меншій кількості. Це забезпечує можливість адекватно оцінити модель на різних сегментах даних.

Розрахуємо модель лінійної регресії. Отримуємо такі дані (Рис 2.4).

Тепер представимо це у вигляді формули. Формула моделі лінійної регресії має такий вигляд:

$$y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \dots + a_n \cdot X_n,$$

де y - залежна змінна, яку ми моделюємо;

$X_1, X_2, \dots, X_n$ - незалежні змінні або фактори;

$a_0, a_1, \dots, a_n$ - коефіцієнти, які були розраховані моделлю лінійної регресії.

В нашому випадку формула буде мати такий остаточний вигляд:

$$Rate = a_0 + a_1 \cdot Votes + a_2 \cdot Duration + a_3 \cdot Cluster2 + a_4 \cdot Cluster3 + a_5 \cdot Cluster4.$$

Підставляючи коефіцієнти:

$$Rate = 5.126 + 7.702 \times 10^{-7} \cdot Votes + 3.822 \times 10^{-3} \cdot Duration + 1.476 \cdot Cluster2 + 1.795 \cdot Cluster3 + 2.362 \cdot Cluster4,$$

де $a_0 - 5.126$ — це константа;

$a_1 - 7.702e-07$ — коефіцієнт для змінної *Votes*. Для кожного додаткового голосу рейтинг фільму збільшується на 7.702×10^{-7} ;

$a_2 - 3.822e-03$ — коефіцієнт для змінної *Duration*. Для кожної додаткової хвилини тривалості фільму рейтинг збільшується на 0.003822;

$a_3 - 1.476$ — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 2 (відносно Кластера 1);

$a_4 - 1.795$ — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 3 (відносно Кластера 1);

$a_5 - 2.362$ — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 4 (відносно Кластера 1).

Residuals:

Min	1Q	Median	3Q	Max
-4.1936	-0.3825	0.0389	0.4513	2.1586

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.126e+00	5.275e-02	97.171	< 2e-16 ***
Votes	7.702e-07	8.117e-08	9.489	< 2e-16 ***
Duration	3.822e-03	5.075e-04	7.532	5.97e-14 ***
Cluster2	1.476e+00	2.559e-02	57.675	< 2e-16 ***
Cluster3	1.795e+00	7.510e-02	23.898	< 2e-16 ***
Cluster4	2.362e+00	3.859e-02	61.205	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6527 on 4657 degrees of freedom

Multiple R-squared: 0.6414, Adjusted R-squared: 0.641

F-statistic: 1666 on 5 and 4657 DF, p-value: < 2.2e-16

Рисунок 2.4 – Результати моделі лінійної регресії

Джерело: розроблено автором.

Кластер 1 використовується як базовий рівень і тому не має власного коефіцієнта у моделі. Вплив кластерів 2, 3 і 4 визначається відносно кластера 1. Формула автоматично враховує кластерну приналежність фільму через коефіцієнти для Cluster2, Cluster3 та Cluster4. Наприклад, якщо фільм належить до Кластера 2, змінна Cluster2 буде дорівнювати 1, а всі інші кластерні змінні будуть дорівнювати 0.

Згідно з результатів можемо побачити, що всі коефіцієнти є значущими у моделі. Кластер 4 має найвищий середній приріст рейтингу на 2,362 пункти порівняно з базовою категорією, що є найсильнішим ефектом серед кластерів. Модель пояснює близько 64,14% варіації рейтингів фільмів, що вважається відносно високим показником, а це означає, що вона є хорошою моделлю для пояснення даних. За значенням F-статистики видно що модель є статистично значущою на загальному рівні.

Тепер перевіримо якість моделі за допомогою метрик, використаних у тестовому наборі даних: RMSE (середньоквадратична похибка) та R^2 . Вони показують частку варіативності, що пояснюється моделлю. Згенеруємо передбачення за допомогою моделі лінійної регресії та візуалізуємо порівняння фактичних значень проти прогнозованих за допомогою графіку розсіювання.

Отримаємо такий графік (Рис. 2.5).

Для більшої наочності я позначив прогнозовані точки зеленим кольором. Як можемо бачити, більшість точок зібрані вздовж і навколо червоної лінії, яка відображає ідеальну відповідність між фактичними і прогнозованими рейтингами. Синьою лінією позначено лінійну регресію. Однак є точки, які віддалені від цієї лінії, що свідчить про помилки в прогнозах. Прогнози здаються більш точними для фільмів із середнім рейтингом близько 5-7, для фільмів з низьким і високим рейтингом прогнози можуть бути не такими точними. Червона лінія вздовж діагоналі графіка - це місце, де були б точки, якби прогнози були ідеальними. Як видно, багато точок знаходяться близько до цієї лінії.

Отже, загальна точність моделі є досить хорошою. На графіку видно, що для низьких фактичних рейтингів, де між 2 і 4, рейтинг моделі буде завищеним. Для

фільмів з дуже високим рейтингом, десь від 8 до 10, прогнози, як правило, будуть заниженими.

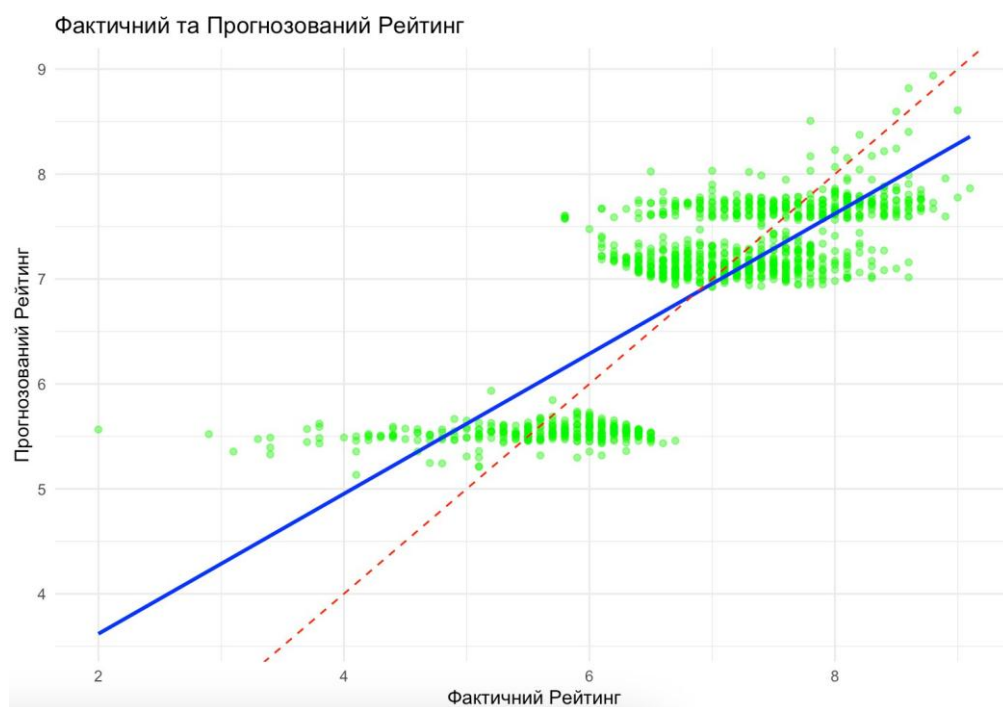


Рисунок 2.5 – Графік розсіювання (Фактичний та Прогнозований рейтинг)

Джерело: складено автором.

Загалом, аналіз рейтингу фільмів IMDb свідчить про те, що різні характеристики фільмів - тривалість, кількість голосів та приналежність до кластеру - мають ключове значення на рейтинги фільмів. При кластерному аналізі фільмів за характеристиками, які мають найбільший вплив на аудиторію, спостерігався чіткий поділ на групи. Це дозволяє виявити закономірності та тренди, характерні для різних груп фільмів. Цілком можливо, що лінійна регресія дозволила добре передбачити фільми в середньому діапазоні, і в той же час її точність знижувалася при спостереженні крайніх значень рейтингів, що свідчить або про необхідність вдосконалення моделі, або про застосування більш складних методів прогнозування, які б більшою мірою враховували різноманітність даних.

Розширивши аналіз, можна виявити, що вдосконалення моделей передбачення рейтингів може бути ключовим для розвитку кіноіндустрії, допомагаючи студіям краще розуміти смаки та вподобання аудиторії та забезпечуючи їм інструменти для створення більш успішних та впізнаваних фільмів.

2.2 Моделювання рейтингу на підґрунті карт Кохонена

Карти Кохонена, або самоорганізуючі карти (Self-Organizing Maps, SOM), є одним з найважливіших інструментів у галузі машинного навчання, особливо коли мова йде про кластеризацію та візуалізацію багатовимірних даних. Ця технологія, розроблена видатним фінським вченим Теуво Кохоненом у 1980-х роках, набула широкого застосування в багатьох сферах науки і техніки завдяки своїй здатності ефективно організовувати та представляти складні дані у зрозумілому вигляді.

Самоорганізуючі карти належать до класу нейронних мереж з некерованим навчанням. Це означає, що вони мають унікальну здатність виявляти приховані структури та закономірності в даних без необхідності попереднього маркування чи надання цільових значень. Такий підхід дозволяє аналізувати великі обсяги інформації, виявляти кластери та подібності між даними, що є надзвичайно корисним у таких галузях, як біоінформатика, фінанси, маркетинг, геологія, та багатьох інших [31].

SOM моделюють спосіб, у який людський мозок обробляє та інтерпретує інформацію, що робить їх потужним інструментом для вирішення складних завдань, пов'язаних з аналізом даних. Вони допомагають зменшити багатовимірність даних, зводячи їх до двовимірного простору, який легко візуалізувати і інтерпретувати. Це забезпечує інтуїтивно зрозумілий спосіб представлення та аналізу інформації, що є ключовим фактором для прийняття обґрунтованих рішень на основі даних.

Метою навчання самоорганізаційної карти (SOM) є навчити різні частини мережі реагувати на певні вхідні шаблони таким чином, щоб забезпечити адекватну кластеризацію та візуалізацію багатовимірних даних. Це досягається шляхом налаштування вагових коефіцієнтів нейронів у мережі таким чином, щоб схожі вхідні дані активували нейрони, розташовані близько один до одного на карті.

Подібний підхід частково мотивується розумінням того, як сенсорна інформація, така як візуальна або слухова, обробляється у відповідних частинах кори головного мозку людини. Наприклад, у зоровій корі є організовані зони, що спеціалізуються на обробці різних аспектів зорової інформації, таких як форма, колір та рух.

Ініціалізація вагових векторів нейронів є важливим кроком у процесі навчання самоорганізаційної карти. Ваги можуть бути ініціалізовані двома способами, а саме:

- **Випадкове ініціалізування:** Ваги нейронів можуть бути встановлені на малі випадкові значення. Це дозволяє мережі почати навчання з різних початкових умов і поступово налаштовуватися під впливом вхідних даних. Випадкове ініціалізування є простим та поширеним методом, який забезпечує необхідне різноманіття для початкового етапу навчання.

- **Ініціалізація на основі головних компонентів:** Ваги нейронів можуть бути ініціалізовані рівномірно з підпростору, охопленого двома найбільшими власними векторами головних компонентів вхідних даних. Цей метод дозволяє значно прискорити процес навчання, оскільки початкові ваги вже дають гарну апроксимацію оптимальних вагових коефіцієнтів. Метод головних компонентів (РСА) зменшує розмірність даних, зберігаючи при цьому найважливіші варіації, що допомагає мережі швидше збігатися до остаточного стану.

Звісно, у кожного методу є свої переваги та недоліки. Наприклад, у методі випадкового ініціалізування з переваг можна відмітити простоту реалізації та забезпечення широкого діапазону початкових умов, що може сприяти більш глобальному охопленню вхідного простору даних. Основні недоліки це повільніша збіжність, оскільки мережа починає працювати з випадковими значеннями і потребує більше часу для налаштування та можливість застрягання в локальних мінімумах. Щодо основних переваг методу ініціалізації на основі головних компонентів, можна виділити швидшу збіжність завдяки кращій початковій апроксимації вагових коефіцієнтів та ефективніше використання інформації про вхідні дані для налаштування мережі. Основні недоліки це те, що цей метод вимагає додаткових обчислень перед початком навчання та є необхідність попереднього

обчислення власних векторів і значень, що може бути обчислювально витратним для великих наборів даних.

Перед початком навчання мережі необхідно зібрати велику кількість вхідних векторів, які представляють різні види даних, очікуваних для відображення. Ці вектори мають бути достатньо різноманітними, щоб охопити всі можливі варіації даних. Ваги нейронів ініціалізуються або малими випадковими значеннями, або рівномірно відбираються з підпростору, охопленого двома найбільшими власними векторами головних компонентів (РСА) вхідних даних. Останній метод дозволяє значно прискорити процес навчання [32].

На кожній ітерації навчання випадковим чином вибирається один вхідний вектор з навчальної множини. Це дозволяє забезпечити рівномірне покриття всіх можливих варіацій даних.

Для обраного вхідного вектора обчислюється евклідова відстань до всіх вагових векторів нейронів мережі. Евклідова відстань визначається як квадратний корінь з суми квадратів різниць відповідних компонентів векторів. Нейрон, чий ваговий вектор найбільш схожий на вхідний вектор (має найменшу евклідову відстань), визначається як найкращий вузол відповідності [33].

Ваги нейронів, розташованих поруч з НВВ у сітці, коригуються таким чином, щоб вони стали ближчими до вхідного вектора. Величина цієї зміни визначається функцією навчання, яка включає коефіцієнт навчання та функцію сусідства. Функція сусідства зменшується з відстанню від НВВ, забезпечуючи, що ближчі нейрони змінюються більше, ніж дальні. З часом коефіцієнт навчання та радіус сусідства зменшуються, що дозволяє системі поступово стабілізуватися. Спочатку зміни ваг є значними, що дозволяє швидко наблизити ваги до вхідних векторів, але з кожною наступною ітерацією ці зміни стають все меншими, що сприяє конвергенції системи.

Описані кроки повторюються для великої кількості вхідних векторів і ітерацій. З кожним циклом навчання ваги мережі стають все ближчими до оптимальних значень, забезпечуючи адекватне відображення багатовимірних даних на двовимірній карті.

Після завершення процесу навчання самоорганізаційна карта здатна ефективно кластеризувати та візуалізувати вхідні дані. Кожен нейрон на карті представляє групу схожих вхідних векторів, що дозволяє легко ідентифікувати схожі та відмінні групи даних. Це робить SOM потужним інструментом для аналізу складних наборів даних, дозволяючи дослідникам та аналітикам знаходити приховані закономірності та тенденції.

Оновлення вагового вектора нейрона v відбувається за наступною формулою:

$$W_v(s + 1) = W_v(s) + \theta(u, v, s) \cdot \alpha(s) \cdot (D(t) - W_v(s)),$$

де s - індекс кроку;

t - індекс в навчальній вибірці;

u - індекс НВВ для вхідного вектора $D(t)$;

$\alpha(s)$ - коефіцієнт навчання, який зменшується з часом;

$\theta(u, v, s)$ - функція сусідства, яка залежить від відстані між нейронами u та v на етапі s .

Функція сусідства $\theta(u, v, s)$ залежить від відстані між найкращим вузлом відповідності (нейрон u) та нейроном v . У найпростішому випадку вона може брати значення 1 для всіх нейронів, достатньо близьких до НВВ, і 0 для інших. Однак зазвичай використовують гауссову функцію. Незалежно від функціональної форми, функція сусідства стискається з часом. Спочатку сусідство є далеким, що дозволяє самоорганізації відбуватися в глобальному масштабі, але з часом воно зменшується до кількох нейронів, дозволяючи точне налаштування ваг [33].

Отже, перейдемо до побудови карт Кохонена. Використовувати будемо вже очищені, стандартизовані дані зі змінними Rate (Рейтинг), Votes (Кількість голосів) та Duration (Тривалість). Також додамо до них вже існуючі кластери. Визначаємо параметри карти Кохонена. Вибираємо 5 нейронів по горизонталі та вертикалі, та обираємо гексагональну топологію, де кожен нейрон має шість сусідів, що забезпечує більш рівномірне покриття і дозволяє краще відобразити структуру даних у багатовимірному просторі. Далі тренуємо карту Кохонена, використовуючи 100 ітерацій та задану швидкість навчання, що дозволить нейронам карти

адаптуватися до структури даних. Мета цього процесу - виявити схожість між спостереженнями та відобразити багатовимірні дані на двовимірній сітці.

Далі виконуємо візуалізацію тренованої карти Кохонена використовуючи різні типи графіків для аналізу та інтерпретації результатів. А саме:

- Візуалізація кодів прототипів;
- Візуалізація змін за епоху;
- Візуалізація кількості спостережень у кожному вузлі;
- Візуалізація дистанцій між вузлами.

Отримаємо такі графіки:

Коди прототипів



Рисунок 2.6 – Коди Прототипів

Джерело: розроблено автором.

Графік "Коди прототипів" (Рис. 2.6) відображає середні значення атрибутів (Rate, Votes, Duration) для кожного нейрона на карті Кохонена. Бачимо, що різні нейрони мають різні комбінації середніх значень атрибутів. Це вказує на те, що карта Кохонена успішно кластеризувала фільми з подібними характеристиками у близькі нейрони. Наприклад, нейрони у верхньому лівому кутку мають високі значення для Рейтингу і низькі для Тривалості, що може вказувати на певний тип фільмів, наприклад короткометражний.

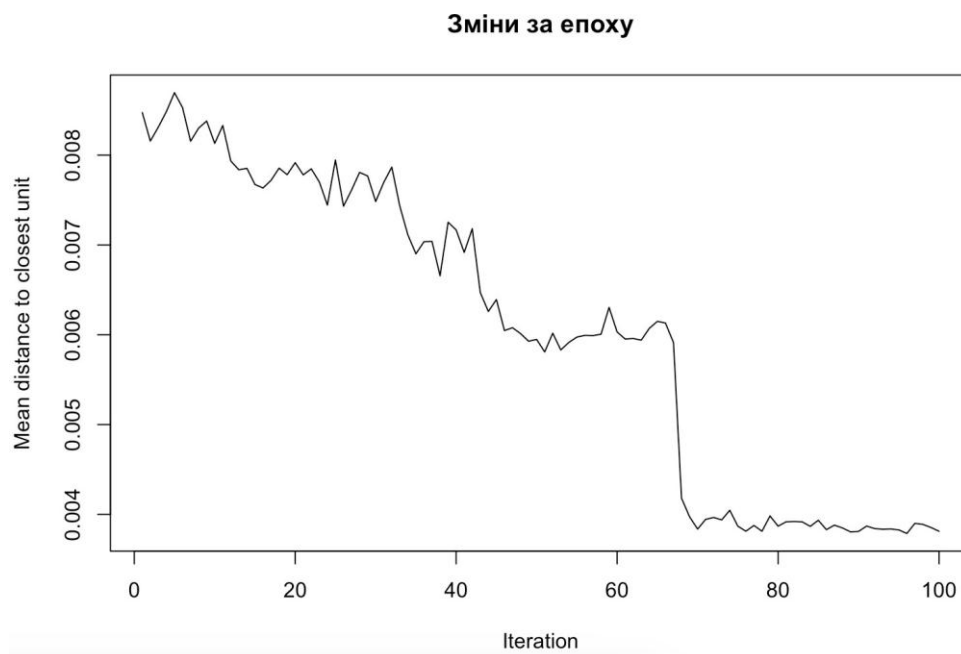


Рисунок 2.7 – Зміни за епоху

Джерело: складено автором.

Графік "Зміни за епоху" (Рис. 2.7) показує, як середня відстань до найближчого нейрона змінюється протягом тренувальних епох. На початку тренування середня відстань значно зменшується, що вказує на швидку адаптацію нейронів до структури даних. Потім зміни стають менш помітними, що свідчить про досягнення стабільного стану.

Кількість спостережень у кожному вузлі

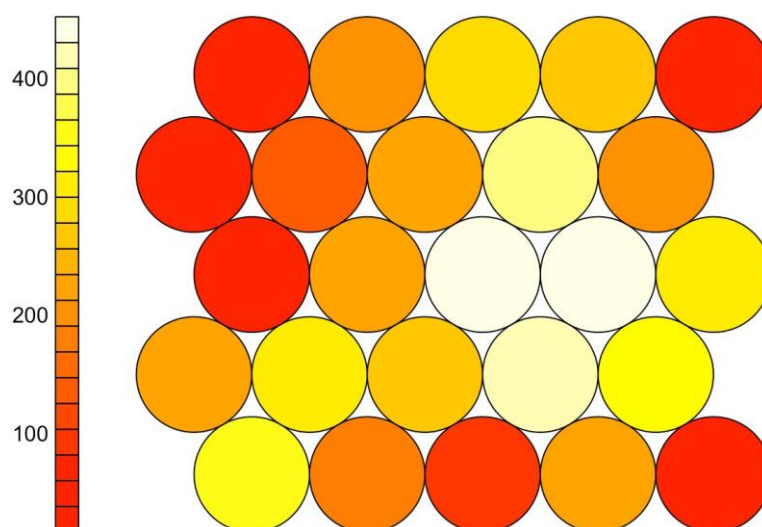


Рисунок 2.8 – Кількість спостережень у кожному вузлі

Джерело: складено автором.

Графік "Кількість спостережень у кожному вузлі" (Рис. 2.8) показує кількість фільмів, що потрапили у кожен нейрон. Колір відображає кількість спостережень: червоний (менше спостережень) та жовтий (більше спостережень). Деякі нейрони містять значно більше спостережень, що може вказувати на існування популярних типів фільмів з подібними характеристиками, тоді як інші нейрони представляють менш поширені типи.

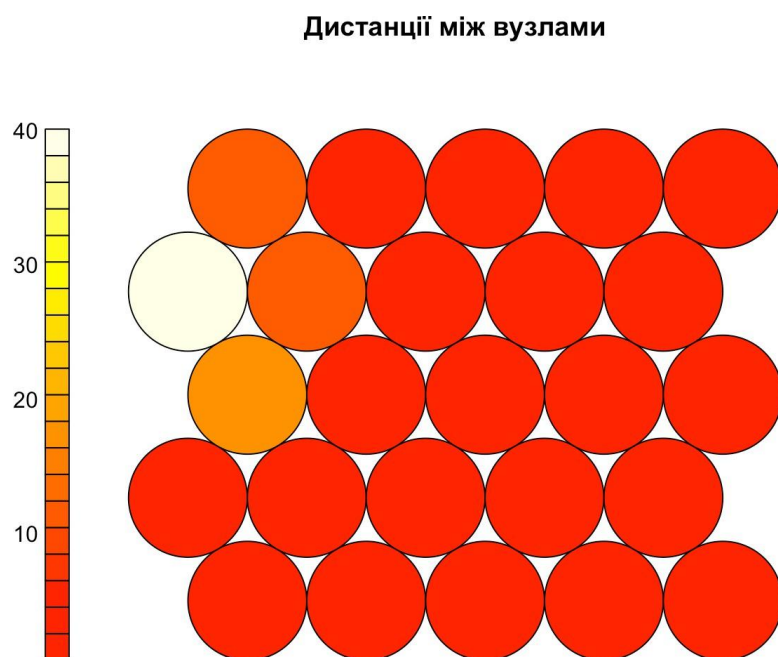


Рисунок 2.9 – Кількість спостережень у кожному вузлі

Джерело: складено автором

Графік "Дистанції між вузлами" (Рис. 2.9) показує дистанції між сусідніми нейронами. Великі дистанції вказані жовтим кольором, тоді як менші дистанції червоним. Великі дистанції між вузлами вказують на межі між різними кластерами. Це дозволяє ідентифікувати основні групи фільмів з подібними характеристиками, які суттєво відрізняються від інших груп. Менші дистанції вказують на те, що сусідні нейрони представляють схожі групи фільмів.

Отже, карти Кохонена успішно кластеризували фільми з подібними характеристиками, що дозволяє візуально ідентифікувати різні групи фільмів. Графіки показують, що модель стабілізувалася після певної кількості епох, і ми можемо чітко побачити межі між основними кластерами фільмів. Це допомагає

зрозуміти структуру даних і визначити основні типи фільмів на основі їх характеристик.

Далі виконаємо додаткову кластеризацію нейронів на карті Кохонена і візуалізуємо ці кластери на самій карті. Ось що ми отримаємо (Рис. 2.10):

Кластери на карті Кохонена

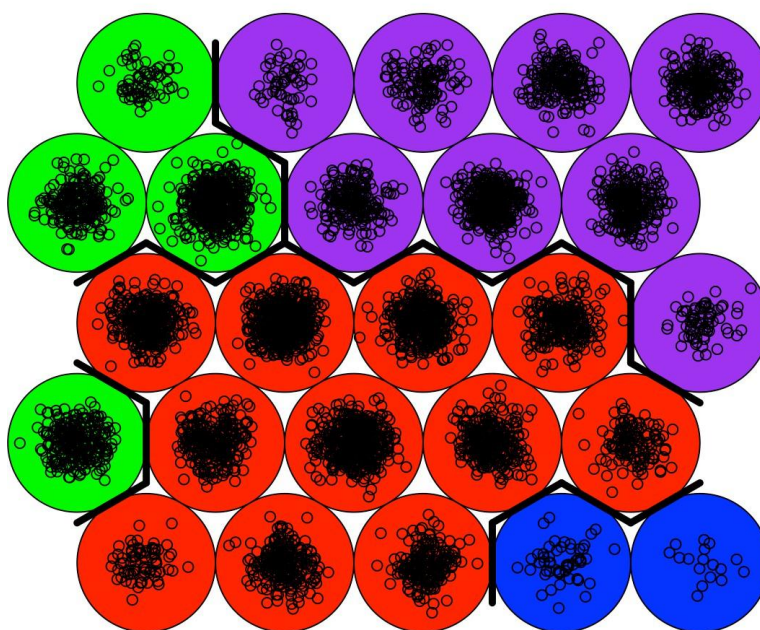


Рисунок 2.10 – Кластери на карті Кохонена

Джерело: розроблено автором.

Для більшої чіткості та зрозумілості я задав колір кластерам. Кластер 1 – червоний колір; Кластер 2 – синій колір; Кластер 3 – зелений колір; Кластер 4 – фіолетовий колір. За результатами можемо сказати, що фільми поділені на чотири основні кластери, що вказує на те, що карта Кохонена успішно виділила основні групи фільмів з подібними характеристиками. Кластеризація показує, що фільми з подібними характеристиками згруповані разом у певних частинах карти. Зелений (3) кластер переважно розташований у верхньому лівому куті. Фіолетовий (4) кластер займає верхню праву частину карти. Червоний (1) кластер займає нижню центральну частину карти. Синій (2) кластер знаходиться у нижньому правому куті. Більшість нейронів мають чіткі кольорові мітки, що вказує на високу щільність спостережень у відповідних кластерах. Червоний (1) кластер є найбільшим за

кількістю нейронів та покриває велику площу на карті, що може вказувати на те, що цей кластер включає в себе більшість фільмів. Інші кластери, такі як синій (2), зелений (3) і фіолетовий (4), також добре визначені і займають окремі частини карти.

Перейдемо до моделювання рейтингу фільмів на підґрунті кластерного аналізу. Розміри вибірок залишаються такими самими як і при кластерному аналізі. Навчальна вибірка містить 4663 спостережень, що є приблизно 80% від загальної кількості записів у датасеті. Містить такі дані, як: Rate (Рейтинг), Votes (Кількість голосів), Duration (Тривалість) та Cluster (Кластер). Тестова вибірка містить приблизно 20% від загальної кількості спостережень, а саме 1164 спостереження.

Розрахуємо модель лінійної регресії. Отримуємо такі дані (Рис. 2.11):

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min       1Q   Median       3Q      Max
-4.2178 -0.3812  0.0399  0.4523  2.1780

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.185e+00  5.189e-02  99.919  < 2e-16 ***
Votes         8.145e-07  8.178e-08   9.959  < 2e-16 ***
Duration      3.324e-03  4.993e-04   6.657  3.12e-11 ***
`as.factor(Cluster)2` 1.467e+00  2.542e-02  57.715  < 2e-16 ***
`as.factor(Cluster)3` 1.752e+00  7.413e-02  23.631  < 2e-16 ***
`as.factor(Cluster)4` 2.322e+00  3.779e-02  61.450  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6448 on 4657 degrees of freedom
Multiple R-squared:  0.6446,    Adjusted R-squared:  0.6442
F-statistic: 1689 on 5 and 4657 DF,  p-value: < 2.2e-16
```

Рисунок 2.11 - Результати моделі лінійної регресії

Джерело: розроблено автором.

В нашому випадку формула моделі лінійної регресії буде мати такий остаточний вигляд:

$$Rate = 5.126 + 8.145 \times 10^{-7} \cdot Votes + 3.324 \times 10^{-3} \cdot Duration + 1.467 \cdot Cluster2 + 1.752 \cdot Cluster3 + 2.322 \cdot Cluster4,$$

де $a_0 = 5.126$ — це константа;

$a_1 - 8.145e-07$ — коефіцієнт для змінної *Votes*. Для кожного додаткового голосу рейтинг фільму збільшується на 8.145×10^{-7} ;

$a_2 - 3.324e-03$ — коефіцієнт для змінної *Duration*. Для кожної додаткової хвилини тривалості фільму рейтинг збільшується на 0.003324;

$a_3 - 1.467$ — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 2 (відносно Кластера 1);

$a_4 - 1.752$ — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 3 (відносно Кластера 1);

$a_5 - 2.322$ — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 4 (відносно Кластера 1).

Згідно з результатами моделі лінійної регресії, всі коефіцієнти є статистично значущими, що підтверджується дуже низькими значеннями *p-value*. Модель лінійної регресії, яка включає кількість голосів (*Votes*), тривалість фільму (*Duration*), та належність до кластера (на основі карт Кохонена), успішно пояснює значну частину варіативності у рейтингах фільмів, а саме 64,46%. Усі предиктори є статистично значущими, що вказує на їх важливість у передбаченні рейтингу. Кластер 4 має найвищий середній приріст рейтингу на 2,362 пункти порівняно з базовою категорією, що є найсильнішим ефектом серед кластерів. Фільми, що належать до різних кластерів, мають різні середні рейтинги, що підкреслює ефективність використання кластерного аналізу для моделювання. Значення *F*-статистики (1689) та його *p-value* ($< 2.2e-16$) свідчать про те, що модель є статистично значущою на загальному рівні, тобто всі предиктори разом взяті значущо впливають на результативну змінну.

Тепер перевіримо якість моделі за допомогою метрик, використаних у тестовому наборі даних: RMSE (середньоквадратична похибка) та R^2 .

Значення $RMSE = 0.661$ свідчать про те, що модель робить точні передбачення з невеликими середніми помилками, а значення $R^2 = 0.627$ вказує на те, що модель добре пояснює варіацію у рейтингах фільмів. Ці результати підтверджують, що використання кластерів на основі карт Кохонена покращує точність моделі і дозволяє ефективніше прогнозувати рейтинги фільмів.

Згенеруємо передбачення за допомогою моделі лінійної регресії та візуалізуємо порівняння фактичних значень проти прогнозованих за допомогою графіку розсіювання.

Отримаємо такий графік (Рис. 2.12):

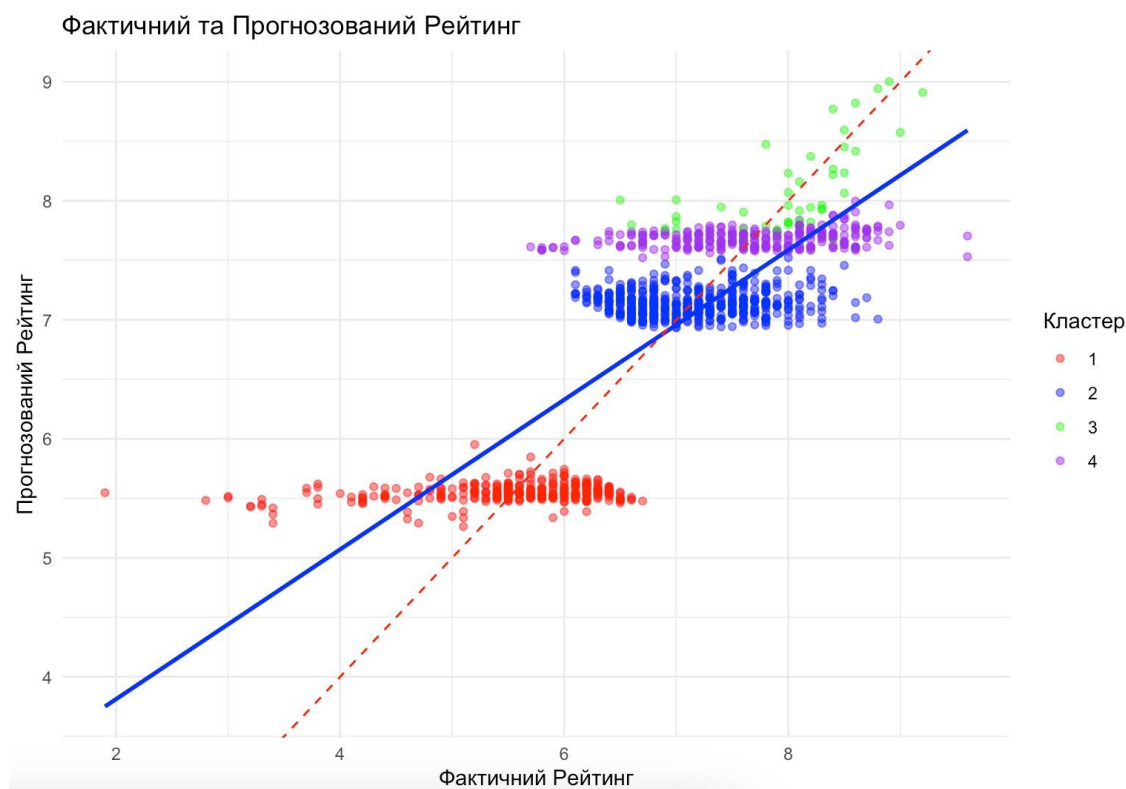


Рисунок 2.12 - Графік розсіювання (Фактичний та Прогнозований рейтинг)

Джерело: розроблено автором.

Згідно з графіком можемо сказати, що більшість точок розташовані близько до червоної лінії, яка символізує ідеальну відповідність між фактичними та прогнозованими рейтингами. Це вказує на те, що модель передбачає значення досить точно. Синьою лінією позначено лінійну регресію. Кластер 1 (червоний): Зосереджений в нижній частині графіку з фактичними рейтингами приблизно від 2 до 6. Кластер 2 (синій): Розташований у середньому діапазоні з фактичними рейтингами приблизно від 5 до 8. Кластер 3 (зелений): Має тенденцію до вищих фактичних рейтингів, приблизно від 6 до 9. Кластер 4 (фіолетовий): Розташований у верхній частині графіку з фактичними рейтингами від 6 до 8.

Деякі точки відхиляються від червоної лінії, що вказує на наявність помилок у передбаченнях. Найбільші відхилення спостерігаються для фільмів з дуже низькими або дуже високими фактичними рейтингами.

2.3 Порівняльний аналіз отриманих результатів та практичні рекомендації

У роботі ми використовуємо два підходи для моделювання рейтингу фільмів: кластерний аналіз та аналіз карт Кохонена. Обидва методи дозволяють групувати фільми за схожими характеристиками та використовувати ці групи для покращення точності прогнозування.

Кластерний аналіз є традиційним методом статистики, який дозволяє виявити природні групи у даних на основі подібності між об'єктами. Цей метод допомагає визначити ключові фактори, що впливають на рейтинги фільмів, та створити сегменти, які мають схожі властивості. Кластерний аналіз є потужним інструментом для виявлення закономірностей та трендів у даних, що дозволяє отримати глибоке розуміння структури кіноіндустрії.

Аналіз карт Кохонена, або самоорганізуючі карти (SOM), представляє інший підхід, заснований на нейронних мережах. Цей метод забезпечує візуалізацію багатовимірних даних у двовимірному просторі, що дозволяє легше інтерпретувати складні дані. SOM здатні адаптуватися до структури вхідних даних, самостійно виявляючи схожі групи без попередньої інформації про класифікацію. Це робить карти Кохонена особливо корисними для дослідження великих обсягів даних та виявлення прихованих закономірностей.

Порівнюючи результати двох підходів, ми оцінюємо їх точність, пояснювальну здатність, переваги та недоліки. Кластерний аналіз надає чіткі та зрозумілі групи, що дозволяє легко інтерпретувати результати та робити конкретні висновки. Проте він може бути обмежений у випадках, коли дані мають складну структуру або невизначені межі між групами. Карти Кохонена, з іншого боку, забезпечують більш гнучкий та інтуїтивно зрозумілий підхід до візуалізації та аналізу даних, але можуть вимагати більше обчислювальних ресурсів та часу для

навчання. Кластерний аналіз може бути особливо корисним для задач, де необхідна чітка сегментація та інтерпретація груп, тоді як карти Кохонена підходять для досліджень з великими та складними даними, де важлива гнучкість та адаптивність методів. Використання обох підходів у комплексі може забезпечити найкращі результати, поєднуючи переваги кожного методу та компенсуючи їх недоліки.

При кластерному аналізі ми використовували метод К-середніх (K-means) для кластеризації фільмів. Цей метод розділяє дані на К груп, мінімізуючи різницю між об'єктами в одному кластері та максимізуючи різницю між об'єктами з різних кластерів. Визначення оптимальної кількості кластерів здійснювалось методом ліктя, який показав, що оптимальною кількістю є 4 кластери.

Формула моделі лінійної регресії має такий вигляд:

$$Rate = 5.126 + 7.702 \times 10^{-7} \cdot Votes + 3.822 \times 10^{-3} \cdot Duration + 1.476 \cdot Cluster2 + 1.795 \cdot Cluster3 + 2.362 \cdot Cluster4,$$

де a_0 - 5.126 — це константа;

a_1 - 7.702e-07 — коефіцієнт для змінної *Votes*. Для кожного додаткового голосу рейтинг фільму збільшується на 7.702×10^{-7} ;

a_2 - 3.822e-03 — коефіцієнт для змінної *Duration*. Для кожної додаткової хвилини тривалості фільму рейтинг збільшується на 0.003822;

a_3 - 1.476 — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 2 (відносно Кластера 1);

a_4 - 1.795 — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 3 (відносно Кластера 1);

a_5 - 2.362 — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 4 (відносно Кластера 1).

При оцінюванні моделі отримали такі результати:

- R^2 : 0.6414 - модель пояснює 64.14% варіації рейтингу. Це значення вказує на те, що модель успішно пояснює більшу частину варіації у рейтингах фільмів. Це досить високе значення для багатьох прикладних задач, що вказує на те, що модель добре справляється з завданням прогнозування рейтингів. Однак, залишаються 35.54% варіації, які модель не пояснює. Це може бути спричинено впливом факторів, які не включені в модель, або випадковими шумами в даних.

Загалом значення можна охарактеризувати як гарне та практично значиме, оскільки модель успішно пояснює більшу частину варіації у рейтингах фільмів і є корисною для прогнозування;

- RMSE: 0.6301 - середньоквадратична похибка. Якщо RMSE значно менше за середнє значення залежної змінної, то модель можна вважати якісною. Наприклад, якщо середній рейтинг фільмів становить 7,0, а RMSE дорівнює 0.6618, то це вказує на те, що модель робить відносно точні прогнози. Конкретно в моєму випадку можна охарактеризувати це значення малим у порівнянні з діапазоном можливих значень рейтингів (1-10). Воно свідчить про відносно низьку середню похибку прогнозів. Отже, модель досить добре передбачає рейтинги фільмів і є якісною у контексті мого аналізу.

Графік розсіювання показує, що модель лінійної регресії, побудована на основі кластерного аналізу, досить добре прогнозує рейтинги фільмів у середньому діапазоні (4-8), але має труднощі з точністю прогнозів для крайніх значень рейтингів. Використання кластерів дозволяє візуально оцінити точність прогнозів для різних груп фільмів, що є корисним інструментом для подальшого вдосконалення моделі.

При аналізі на основі карт Кохонена ми використовували сітку розміром 5x5 з гексагональними вузлами. Модель була натренована на 100 епохах з початковою швидкістю навчання 0.05, яка знижувалася до 0.01.

Формула моделі лінійної регресії має такий вигляд:

$$Rate = 5.126 + 8.145 \times 10^{-7} \cdot Votes + 3.324 \times 10^{-3} \cdot Duration + 1.467 \cdot Cluster2 + 1.752 \cdot Cluster3 + 2.322 \cdot Cluster4,$$

де a_0 - 5.126 — це константа;

a_1 - 8.145e-07 — коефіцієнт для змінної *Votes*. Для кожного додаткового голосу рейтинг фільму збільшується на 8.145×10^{-7} ;

a_2 - 3.324e-03 — коефіцієнт для змінної *Duration*. Для кожної додаткової хвилини тривалості фільму рейтинг збільшується на 0.003324;

a_3 - 1.467 — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 2 (відносно Кластера 1);

$a_4 - 1.752$ — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 3 (відносно Кластера 1);

$a_5 - 2.322$ — додатковий приріст до рейтингу для фільмів, що належать до Кластеру 4 (відносно Кластера 1).

При оцінюванні моделі отримали такі результати:

- R^2 : 0.6446 - модель пояснює 64.46% варіації рейтингу. Це значення вказує на те, що модель досить добре справляється з передбаченням рейтингів, хоча ще залишається простір для покращення;
- RMSE: 0.6618 - середньоквадратична похибка. Для рейтингових даних, значення RMSE близько 0.66 вважається прийнятним. Це означає, що модель робить досить точні передбачення, хоча існують деякі помилки.

Графік розсіювання показує, що модель лінійної регресії, побудована на основі карт Кохонена, досить добре прогнозує рейтинги фільмів у середньому діапазоні (5-8), але має труднощі з точністю прогнозів для крайніх значень рейтингів. Кластери, відображені різними кольорами, дозволяють візуально оцінити точність прогнозів для різних груп фільмів. Особливо добре модель передбачає рейтинги для Кластера 2 (синій) та Кластера 3 (зелений), тоді як для Кластера 1 (червоний) і Кластера 4 (фіолетовий) спостерігаються більші відхилення від червоної лінії. Це свідчить про те, що використання карт Кохонена допомагає покращити загальну точність прогнозів, але є простір для подальшого вдосконалення моделі, особливо для фільмів з дуже низькими або дуже високими рейтингами.

Перейдемо до порівняльної оцінки точності обох моделей.

Коефіцієнт детермінації (R^2): Кластерний аналіз має значення 0.6414, що вказує на те, що модель пояснює 64.14% варіації залежної змінної. Це досить високий показник, що свідчить про добру якість моделі. Карти Кохонена мають трохи вище значення коефіцієнта детермінації 0.6446, що вказує на те, що ця модель пояснює 64.46% варіації залежної змінної. Це означає, що обидві моделі здатні пояснити понад 64% варіації залежної змінної. Однак, модель на основі карт

Кохонена показує трохи кращий результат, що свідчить про її трохи вищу пояснювальну здатність.

Середньоквадратична похибка (RMSE): RMSE вимірює середню величину помилки між прогнозованими та фактичними значеннями. Кластерний аналіз з RMSE 0.6301 показує, що середня похибка прогнозів цієї моделі є меншою, що робить її більш точною у передбаченні результатів у порівнянні з моделлю на основі карт Кохонена. Карти Кохонена мають RMSE 0.6618, що вказує на трохи більшу середню похибку. Хоча різниця незначна, кластерний аналіз має перевагу в точності прогнозів.

Значення F-статистики: F-статистика використовується для перевірки загальної значущості моделі. Вона порівнює пояснювану варіацію з непояснюваною. Кластерний аналіз має значення F-статистики 1666, що свідчить про високу значущість моделі. Це означає, що модель добре пояснює варіації даних. Карти Кохонена з значенням F-статистики 1689 мають трохи вищу значущість. Це підтверджує, що ця модель є трохи кращою у загальному поясненні варіацій даних.

При виборі між цими двома моделями важливо враховувати конкретні цілі аналізу. Якщо пріоритетом є пояснювальна здатність моделі, то карти Кохонена мають невелику перевагу. Вони пояснюють трохи більшу частку варіації залежної змінної ($R^2 = 0.6446$), що може бути важливим при аналізі даних для розуміння основних закономірностей і взаємозв'язків.

Якщо ж важлива точність прогнозів, то краще вибрати кластерний аналіз. З меншою середньоквадратичною похибкою (RMSE = 0.6301), ця модель забезпечує точніші прогнози, що може бути критично важливим у ситуаціях, де необхідно мінімізувати похибки, наприклад, у фінансовому прогнозуванні або управлінні ризиками. Для загальної значущості моделі карти Кохонена мають невелику перевагу через трохи вищу F-статистику (1689). Це свідчить про те, що моделі на основі карт Кохонена можуть краще справлятися із загальним поясненням варіацій даних.

Отже, будь-якій з двох моделей може бути відданий пріоритет. Що вже залежить від потреб. Для кращої пояснювальної здатності і загальної значущості,

карти Кохонена є трохи кращими. Однак, якщо важлива точність прогнозів, то кластерний аналіз є кращим вибором. У будь-якому випадку, обидві моделі є досить потужними і можуть бути корисними у різних контекстах аналізу даних.

Результати моделювання можуть бути інтегровані у рекомендаційні системи для онлайн-стримінгових платформ, таких як Netflix або Amazon Prime. Ці моделі допоможуть передбачити рейтинг фільмів для конкретних користувачів, враховуючи різні параметри, такі як тривалість фільму, кількість голосів та належність до певного кластеру. Завдяки використанню кластерного аналізу можна групувати фільми за схожими характеристиками, що дозволяє розробити точніші алгоритми для прогнозування вподобань користувачів. Наприклад, фільми з подібним жанром, тривалістю або акторським складом можуть бути віднесені до одного кластеру, і на основі цього можна створювати рекомендації для користувачів з відповідними інтересами.

Використання карт Кохонена може бути особливо корисним для ідентифікації більш точних груп користувачів з подібними уподобаннями. Карти Кохонена дозволяють візуалізувати багатовимірні дані та знаходити приховані закономірності, що допомагає краще розуміти структуру даних про користувачів. Це дозволяє створювати більш персоналізовані рекомендації, враховуючи індивідуальні особливості кожного користувача.

Продюсери та кінокомпанії можуть використовувати результати кластерного аналізу та аналізу карт Кохонена для прийняття рішень щодо виробництва нових фільмів. Аналізуючи, які кластери фільмів мають найвищі рейтинги та найбільшу кількість голосів, компанії можуть зосередити свої зусилля на створенні контенту, що відповідає успішним кластерам. Наприклад, якщо аналіз показує, що фільми з певного жанру або з участю конкретних акторів отримують високі оцінки, кінокомпанії можуть інвестувати в проекти з подібними характеристиками. Карты Кохонена допомагають виявити приховані зв'язки між різними характеристиками фільмів та уподобаннями аудиторії. Це дозволяє продюсерам точніше прогнозувати, які типи фільмів будуть користуватися популярністю. В результаті,

компанії можуть знижувати ризики, пов'язані з виробництвом нових фільмів, і зосереджувати свої ресурси на проектах з високим потенціалом успіху.

Також результати кластерного аналізу та карт Кохонена можна використовувати для маркетингових цілей, наприклад розробки таргетованих кампаній. Наприклад, фільми, що належать до кластерів з високим рейтингом, можуть бути просунуті серед користувачів, які мають подібні уподобання, визначені через аналіз карт Кохонена. Це дозволить підвищити ефективність маркетингових зусиль, оскільки реклама буде націлена на аудиторію, яка з найбільшою ймовірністю зацікавиться пропонованим контентом.

Розширення моделей за рахунок включення додаткових змінних може значно покращити точність прогнозування рейтингів фільмів. З важливих змінних можна відмітити такі: жанр, акторський склад, режисер, бюджет фільму. Інтеграція цих змінних у моделі дозволить врахувати більшу кількість факторів, що впливають на уподобання глядачів, тим самим підвищуючи точність прогнозування рейтингів.

Для покращення точності прогнозів можна розглянути використання більш складних моделей машинного навчання, таких як метод випадкових лісів, нейронні мережі, градієнтний бустинг. Ці алгоритми можуть краще враховувати нелінійні взаємозв'язки між змінними та забезпечити вищу точність прогнозів у порівнянні з традиційними методами. Також для покращення точності прогнозів важливо враховувати фактори, що відображають сезонні коливання популярності фільмів, а також довгострокові тренди. Наприклад, деякі жанри можуть бути популярнішими в певні періоди року, такі як свята або літні канікули. Включення цих факторів у моделі дозволить точніше прогнозувати рейтинги в залежності від часу випуску фільму. Та звичайно регулярне оновлення моделей з урахуванням нових даних також є критично важливим для збереження їх точності. З часом смаки та уподобання аудиторії можуть змінюватися, тому важливо постійно оновлювати дані та перенавчати моделі. Це дозволить підтримувати актуальність моделей і забезпечити їх високу продуктивність у довгостроковій перспективі.

Отже, результати моделювання рейтингів фільмів на основі кластерного аналізу та аналізу карт Кохонена можуть мати широкий спектр застосувань у

кіноіндустрії та стримінгових сервісах. Використання цих моделей дозволяє краще розуміти уподобання глядачів, оптимізувати процес виробництва та просування фільмів, а також покращувати рекомендаційні системи. Для подальшого покращення моделей варто враховувати додаткові фактори, використовувати складніші алгоритми та регулярно оновлювати моделі на основі нових даних. Це забезпечить високу точність та надійність прогнозів, сприяючи успіху кінопроектів та задоволенню потреб аудиторії.

ВИСНОВКИ

У сучасному середовищі кіноіндустрії стрімінгові сервіси та цифрові платформи стали дуже популярними. Це сталося головним чином тому, що такі сервіси, як Netflix, Amazon Prime Video, HBO, Hulu та інші, дали можливість глядачам отримати доступ до перегляду фільмів і серіалів у будь-який час і з будь-якого місця. Фактично, це стало основою для прискорення потоку контенту до споживачів і ще більше прискорило конкуренцію між кінематографістами та телепродюсерами.

Зростання конкуренції в кіноіндустрії спонукає до пошуку нових способів просування фільмів для залучення до них аудиторії. Водночас рейтинги фільмів у сучасній ринковій системі, що характеризується насиченістю контенту, стають критично важливим процесом на шляху до комерційної вигоди. Вони допомагають визначити найкращі та найпопулярніші серед великої кількості контенту, а також полегшують глядачам вибір та допомагають їм у прийнятті рішень.

Рейтинги фільмів на IMDb та інших подібних сайтах займають важливе місце у формуванні індивідуальних рішень глядачів. Вони слугують своєрідним бенчмарком, який дозволяє швидко оцінити якість та популярність фільму за допомогою рейтингу та відгуків інших людей. Це може відбуватися з тієї причини, що в умовах обмеженого часу глядач відчуває бажання зробити правильний вибір, не витрачаючи надто багато часу на перегляд безлічі трейлерів чи читання рецензій. Вони не лише впливають на індивідуальні рішення глядачів, але й формують загальні тенденції та вподобання у світі кіно. Високі рейтинги можуть збільшити популярність фільму, залучити до нього більше глядачів та навіть вплинути на його комерційний успіх. У свою чергу, низькі рейтинги можуть негативно позначитися на переглядах та прибутках. Таким чином, рейтинги виступають як індикатор якості та популярності фільмів, що впливає на поведінку глядачів та тенденції ринку. Це дуже важливо для розуміння механізмів, що лежать в основі формування рейтингів у кіноіндустрії. Це дослідження допомагає виявити ключові фактори, що

впливають на оцінки користувачів, та сформувати стратегію покращення рейтингу фільмів.

Для досягнення мети дослідження мною були успішно виконані усі поставлені завдання роботи. Проведено глибокий аналіз сутності рейтингового управління в економіці, враховуючи його особливості та вплив на зовнішні чинники. Вивчено специфіку кіноіндустрії та обґрунтовано необхідність рейтингування фільмів у цьому контексті. Розглянуто платформу IMDb як найбільш відомий та використовуваний ресурс для рейтингування фільмів, зокрема, проведено огляд чинників, що впливають на формування рейтингу на цій платформі. Застосовані сучасні методи аналізу даних, зокрема кластерний аналіз та самоорганізуючі карти Кохонена. Обидва підходи продемонстрували високу ефективність у групуванні та прогнозуванні рейтингів фільмів, дозволяючи отримати цінні інсайти для подальшого використання в кіноіндустрії та на стримінгових платформах.

Кластерний аналіз, заснований на методі К-середніх, показав, що фільми можуть бути розподілені на чотири основні кластери, що відрізняються за рейтингом, кількістю голосів та тривалістю. Модель лінійної регресії, побудована на основі кластерного аналізу, забезпечила високу точність передбачень, пояснюючи більше 64% варіації рейтингів фільмів. Це підтверджує, що кластерний аналіз є потужним інструментом для виявлення закономірностей у великих масивах даних.

Модель лінійної регресії, побудована на основі карт Кохонена, також показала високий рівень точності, підтверджуючи ефективність цього методу для аналізу даних та виявлення закономірностей. Використання карт Кохонена дозволило більш глибоко зрозуміти структуру даних та побачити приховані взаємозв'язки. Загалом, проведене дослідження показало, що поєднання кластерного аналізу та карт Кохонена може забезпечити найкращі результати для моделювання рейтингів фільмів.

Рекомендації, надані у роботі, передбачають впровадження нових технологій та підходів до аналізу даних, що сприятиме подальшому розвитку

кіноіндустрії. Використання інноваційних методів аналізу, таких як кластерний аналіз та самоорганізуючі карти Кохонена дозволяє отримати більш глибоке та точне розуміння вподобань аудиторії. Це відкриває нові можливості для створення контенту, який максимально відповідає потребам і очікуванням глядачів. Завдяки цим підходам, кіностудії та продюсери можуть точніше визначати потенційну аудиторію для нових фільмів, ефективніше планувати рекламні кампанії та розробляти стратегії випуску, які сприятимуть підвищенню інтересу до фільму. Це, у свою чергу, сприятиме підвищенню задоволеності глядачів та збільшенню прибутків від кінофільмів. Крім того, такі підходи дозволяють виявляти нові тренди та вподобання аудиторії, що є надзвичайно важливим для довгострокового планування та успіху у кіноіндустрії. Також виконано порівняльний аналіз моделей моделювання рейтингу фільмів на прикладі IMDb. На основі результатів дослідження сформульовано висновки та надано практичні рекомендації щодо застосування моделей рейтингування у практичній діяльності в кіноіндустрії.

Тож, використання сучасних методів аналізу даних не лише підвищує ефективність прогнозування та управління рейтингами фільмів, але й сприяє створенню якіснішого та більш привабливого контенту, що відповідає очікуванням сучасних глядачів. Це є ключовим фактором для підтримання конкурентоспроможності та успіху у динамічному середовищі сучасної кіноіндустрії.

Результати цієї роботи відкривають нові перспективи для подальших досліджень у галузі аналізу та моделювання рейтингів. Майбутні дослідження можуть зосередитися на вдосконаленні існуючих методів, розробці нових підходів та їхньому практичному застосуванні для більш ефективного прогнозування та управління рейтингами фільмів. Наприклад, можна досліджувати інтеграцію додаткових даних, таких як соціальні медіа, відгуки критиків та глядацькі тренди, для підвищення точності прогнозів. Подальші дослідження можуть також сприяти розвитку автоматизованих систем рекомендацій, що допомагають глядачам знаходити фільми, які найкраще відповідають їхнім вподобанням. Це може включати розробку алгоритмів машинного навчання та штучного інтелекту, які

аналізують великі масиви даних і надають персоналізовані рекомендації. Крім того, вивчення впливу культурних, демографічних та регіональних факторів на рейтинги може дати цінні інсайти для створення контенту, який відповідає очікуванням різних аудиторій по всьому світу.

Отже, дослідження формування та моделювання рейтингів фільмів має велике значення як для теоретичних, так і для практичних аспектів кіноіндустрії, сприяючи її постійному розвитку та вдосконаленню.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Миронова Л.Г. Аналіз досвіду рейтингового оцінювання і управління в сучасній економіці. *Інвестиції: практика та досвід*, 2017. № 17. С. 40-43.
2. Боднарук І.Л., Шлапак Л.С. Особливості етапів формування рейтингових оцінок у системі рейтингового управління. *Науковий Вісник НГУ*, 2018. С. 136-141.
3. Lahti A. Foundations Of The Revolutionary Rating Method For Innovative Smes. *Journal of Enterprising Culture (JEC)*. World Scientific Publishing Co. Pte. Ltd. vol. 6(02), 1998. P. 171-197.
4. Bar-Isaac H., Shapiro J. D. Ratings Quality Over the Business Cycle : NYU and University of Oxford, 2010. 38 p.
5. Gavilan D., Avello M., Martinez-Navarro G. The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, 2018. vol. 66. P. 53-61.
6. Bargagliotti A., Lingfang L. Decision Making Using Rating Systems: When Scale Meets Binary, 2012. 21 p.
7. Fisher M., Newman G., Dhar R. Seeing Stars: How the Binary Bias Distorts the Interpretation of Customer Ratings. *Journal of Consumer Research*. vol. 45(03), 2018. P. 471-489.
8. Проскуріна М.О. Кіноіндустрія як сфера економічної діяльності. *Економіка та управління національним господарством*, 2018. № 15. С. 23-29.
9. Kerrigan F., Ngan E., Shannon R. Entrepreneurship and marketing in the film industry. *Handbook of Entrepreneurship and Marketing*, 2020. P. 305–319.
10. Rathnayaka S. J. J., Ranathunga C. J., Navarathna R., Kaneswaran A., Balathanan Y. *Predicting Movie Ratings from Audience Behaviors on Movie Trailers. 10th International Conference on Information and Automation for Sustainability (ICIAfS)*. Sri Lanka, 2021. P. 489-493.
11. Миславський В.Н. Кінословник. Терміни, визначення, жаргонізми. Харків, 2007. 328 с.

12. Ramos M., Calvão A., Anteneodo C. Statistical Patterns in Movie Rating Behavior. *PLOS ONE*, 2015. 17 p.
13. Moon, S., Bergey P., Iacobucci D. Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction. *Journal of Marketing*, 2010. vol. 74. P. 108-121.
14. Palsson C., Price J., Shores J. Ratings and Revenues: Evidence from Movie Ratings. *Contemporary Economic Policy*, 2013. P. 3-21.
15. Gentile D. Are motion picture ratings reliable and valid?. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, 2010. vol. 47(05). P. 423-424.
16. Knight S. Col Needham of IMDb: 'Everything about me is movies'. *Financial Times*. URL: <https://www.ft.com/content/01f1230c-afcb-11e3-9cd1-00144feab7de#axzz2whIXmQ2w> . (дата звернення: 03.05.2024).
17. Lewis R. IMDb. URL: <https://www.britannica.com/topic/IMDb> . (дата звернення: 03.05.2024).
18. What's included with an IMDbPro membership? URL: <https://help.imdb.com/article/imdbpro/membership-benefits/what-s-included-with-an-imdbpro-membership/G6EFQ2AYEG3NKTm2#>. (дата звернення: 04.05.2024).
19. imdb.com Traffic & Engagement Analysis. URL: <https://www.similarweb.com/website/imdb.com/#traffic> . (дата звернення: 04.05.2024).
20. IMDb Statistics. URL: <https://www.imdb.com/pressroom/stats/> (дата звернення: 04.05.2024).
21. Рейтинг IMDb Top 250 – Как он составляется?. URL: <https://kinowar.com/rejtyng-imdb-top-250-kak-on-sostavlyaetsya/> . (дата звернення: 04.05.2024).
22. Oghina A., Breuss M., Tsagkias M., Rijke M. Predicting IMDB Movie Ratings Using Social Media, 2012. P. 503-507.

23. Saraee M., White S., Eccleston J. A data mining approach to analysis and prediction of movie ratings. *Data Mining V*, 2004. vol. 33. P. 343-352.
24. Zhipeng J., Qiudan L., Daniel Dajun Z., Yongcheng Z., Ruoran L., Lei W., Hongyuan M. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). *KDD '14: Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*: New York, August 24-27, 2014. P. 193-202.
25. Çizmecı B., Ögüdücü Ş. G. Predicting IMDb Ratings of Pre-release Movies with Factorization Machines Using Social Media. *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. Sarajevo, Bosnia and Herzegovina, 2018. P. 173-178.
26. Marovic M., Mihokovic M., Miksa M., Pribil S., Tus A. Automatic movie ratings prediction using machine learning. *MIPRO, 2011 Proceedings of the 34th International Convention*, Opatija, Croatia, 23-27 May, 2011. P. 1640-1645.
27. Everitt B. Cluster analysis. *Quality & Quantity: International Journal of Methodology*, 1980. vol.14(01). P. 75-100.
28. Johnson R. A., Wichern D. W. Applied Multivariate Statistical Analysis. Fifth Edition, 2002. 767 p.
29. Kaufman L., Rousseeuw P. J. Finding Groups in Data: An Introduction to Cluster Analysis, 1990. 342 p.
30. Jain A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 2010. vol. 31(08). P. 651-666.
31. Hastie T., Tibshirani R., Friedman J. Self-Organizing Maps. *The Elements of Statistical Learning*, 2009. P. 528-534.
32. Flanagan J.A. Self-organisation in Kohonen's SOM. *Neural networks : the official journal of the International Neural Network Society*, 1996. vol. 9(07). P. 1185-1197.
33. Vishwanathan S., Murty M. Kohonen's SOM with cache. *Pattern Recognition*, 2000. vol.33(11). P. 1927-1929.

ДОДАТКИ

Додаток А

Таблиця А.1 – Структура датасету, використаного в проведенні дослідження

Name	Date	Rate	Votes	Genre	Duration	Type	Certificate	Episodes	Nudity	Violence	Profanity	Alcohol	Frightening
No Time to Die	2021	7,6	107163	Action, Adventure	163	Film	PG-13	-	Mild	Moderate	Mild	Mild	Moderate
The Guilty	2021	6,3	64375	Crime, Drama, Thriller	90	Film	R	-	None	None	Severe	None	Moderate
The Many Saints of	2021	6,4	27145	Crime, Drama	120	Film	R	-	Moderate	Severe	Severe	Moderate	Moderate
Venom: Let There Be Carnage	2021	6,4	30443	Action, Adventure	97	Film	PG-13	-	None	Moderate	Moderate	Mild	Moderate
Dune	2021	8,3	84636	Action, Adventure	155	Film	PG-13	-	None	Moderate	None	Mild	Moderate
Ted Lasso	2021	8,8	117562	Comedy, Drama, Sports	30	Series	TV-MA	23	Mild	None	Severe	Mild	None
Free Guy	2021	7,3	153835	Action, Adventure	115	Film	PG-13	-	Mild	Moderate	Moderate	None	Mild
House of the Dragon	2021			Action, Adventure, Drama		Series	None	10	No Rate	No Rate	No Rate	No Rate	No Rate
What If...?	2021	7,6	51959	Animation, Action, Adventure		Series	TV-14	18	None	Moderate	Mild	Mild	Moderate
Sex Education	2021	8,3	220201	Comedy, Drama	45	Series	TV-MA	25	Severe	Mild	Severe	Moderate	Mild
The Walking Dead	2021	8,2	906327	Drama, Horror, Thriller	44	Series	TV-14	177	Mild	Severe	Moderate	Mild	Severe
The Chestnut	2021	7,9	15434	Crime, Drama, Mystery	50	Series	TV-MA	6	Mild	Moderate	Mild	None	Moderate
Foundation	2021	7,4	16675	Drama, Sci-Fi		Series	TV-14	11	Mild	Moderate	Mild	No Rate	Mild
The Sopranos	2021	9,2	340006	Crime, Drama	55	Series	TV-MA	86	Severe	Severe	Severe	Severe	Severe
Only Murders in the Street	2021	8,3	22988	Comedy, Crime, Drama		Series	TV-MA	11	Mild	Moderate	Severe	Mild	Mild
American Horror Story	2021	8,0	296306	Drama, Horror, Thriller	60	Series	TV-MA	116	Severe	Severe	Severe	Severe	Severe
Seinfeld	2021	8,8	272028	Comedy	22	Series	TV-PG	174	Mild	Mild	Mild	Mild	None
Grey's Anatomy	2021	7,5	275094	Drama, Romance	41	Series	TV-14	384	Moderate	Severe	Mild	Moderate	Moderate
Game of Thrones	2021	9,2	1885776	Action, Adventure	57	Series	TV-MA	73	Severe	Severe	Severe	Moderate	Severe
Black Widow	2021	6,8	246603	Action, Adventure	134	Film	PG-13	-	None	Moderate	Mild	Mild	Moderate
The Last Duel	2021	7,7	6498	Action, Drama, History	152	Film	R	-	Severe	Severe	Moderate	Mild	Severe
Alice in Borderland	2021	7,7	33151	Action, Fantasy, Mystery	50	Series	TV-MA	9	Moderate	Severe	Mild	Mild	Severe
Shang-Chi and the Legend of the Ten Rings	2021	7,9	109552	Action, Adventure	132	Film	PG-13	-	None	Moderate	Mild	Mild	Mild
Billions	2021	8,4	80562	Drama	60	Series	TV-MA	72	Moderate	None	Severe	Moderate	None
Resident Evil	2021			Action, Horror, Mystery	107	Film	R	-	No Rate	No Rate	No Rate	No Rate	No Rate
Spectre	2015	6,8	407514	Action, Adventure	148	Film	PG-13	-	Mild	Moderate	Mild	Mild	Moderate
Casino Royale	2006	8,0	613873	Action, Adventure	144	Film	PG-13	-	Mild	Moderate	Mild	Mild	Moderate
Lucifer	2006	8,1	288801	Crime, Drama, Fantasy	42	Series	TV-14	93	Mild	Mild	Mild	Moderate	Mild
The Morning Show	2006	8,4	73736	Drama	60	Series	TV-MA	20	Mild	None	Severe	Mild	Mild
Halloween	2021	6,1	20383	Horror, Thriller	105	Film	R	-	None	Severe	Severe	Mild	Severe
There's Something in the Water	2021	4,8	7724	Horror, Mystery, Thriller	96	Film	TV-MA	-	Mild	Severe	Moderate	Moderate	Moderate
Malignant	2021	6,3	5331	Crime, Horror, Mystery	111	Film	R	-	None	Severe	Moderate	None	Severe
Wolf	2021	4,1	229	Drama, Mystery, Thriller	98	Film	None	-	No Rate	No Rate	No Rate	No Rate	No Rate
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Джерело: сайт <https://www.kaggle.com>.