6. A Hybrid Deep Learning and Bayesian Network Model for Decision-Making under Uncertainty / S. Zhu, B. Chen, S. Gao et al. *2021 IEEE International Conference on Prognostics and Health Management (ICPHM).* URL: https://ieeexplore.ieee.org/document/9486527

7. Barber D. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

УДК 004.82:004.85]:791.6(043.2)

Левченко М. Р., *здобувачка вищої освіти 3 курсу спеціальності 122 Комп'ютерні науки,*
Хмелівський Ю. С., *асистент кафедри інформаційних технологій*

АНАЛІЗ ХАРАКТЕРИСТИК ФІЛЬМІВ ТА ЇХ УСПІШНОСТІ ЗА ДОПОМОГОЮ DATASET MOVIES У МОВІ R

Донецький національний університет імені Василя Стуса, м. Вінниця

Аналіз даних є важливим інструментом для виявлення закономірностей у великих масивах інформації. У цій роботі досліджується набір даних про фільми (movies) з пакету ggplot2movies [1] у середовищі R, який містить інформацію про жанри, рік випуску, тривалість, рейтинг та інші характеристики понад 58 тисяч фільмів. Метою дослідження є вивчення структури даних за допомогою кластерного аналізу та побудова моделей класифікації для прогнозування успішності фільмів. Робота базується на методах ієрархічної кластеризації, K-середніх та логістичної регресії, а також демонструє можливості мови R для статистичного аналізу та візуалізації.

Почнемо з кластерного аналізу [2].

Лістинг програми:

```
library(ggplot2movies)
library(dplyr)
data("movies")
movies_clean <- movies %>%
  select(title, year, length, rating, votes, Action, Comedy, Drama, Romance,
Short) %>%
  na.omit()
summary(movies_clean[, c("year", "length", "rating", "votes")])
colMeans(movies_clean[, c("Action", "Comedy", "Drama", "Romance", "Short")])
```

Результат:

Після очистки залишилось 58 788 фільмів і 10 змінних. Цільової змінної немає, тому задача кластеризації є без учителя. У більшості фільмів тривалість приблизно 82 хв, рейтинг – 5.9, а медіана голосів – 30. Найпоширеніші жанри: драма (37 %), комедія (29 %) та короткометражки (16 %).

```
Кількість об'єктів у вибірці: 58788
> cat("Кількість атрибутів:", ncol(movies_clean), "\n")
Кількість атрибутів: 10
>
> cat("Цільовий атрибут відсутній – це задача кластерного аналізу.\n")
Цільовий атрибут відсутній – це задача кластерного аналізу.
>
> summary(movies_clean[, c("year", "length", "rating", "votes")])
      year          length            rating
 Min.   :1893   Min.   :   1.00   Min.   : 1.000
 1st Qu.:1958   1st Qu.:  74.00   1st Qu.: 5.000
 Median :1983   Median :  90.00   Median : 6.100
 Mean   :1976   Mean   :  82.34   Mean   : 5.933
 3rd Qu.:1997   3rd Qu.: 100.00   3rd Qu.: 7.000
 Max.   :2005   Max.   :5220.00   Max.   :10.000
     votes
 Min.   :     5.0
 1st Qu.:    11.0
 Median :    30.0
 Mean   :   632.1
 3rd Qu.:   112.0
 Max.   :157608.0
>
> genre_props <- colMeans(movies_clean[, c("Action", "Comedy", "Drama",
mance", "Short")])
> print(round(genre_props, 3))
 Action  Comedy   Drama Romance   Short
  0.080   0.294   0.371   0.081   0.161
```

Рис. 1. Результат кластерного аналізу

Далі було проведено кластеризацію. Для ієрархічної взято випадкову підвибірку з 1 000 фільмів, щоб уникнути перевантаження пам’яті. За результатами, оптимальною виявилась кількість кластерів – чотири [3].

Лістинг програми:

```
library(factoextra)
library(dplyr)
set.seed(123)
data_cluster <- movies_clean %>%
  select(length, rating, votes, Action, Comedy, Drama, Romance, Short) %>%
  scale()
sample_data <- data_cluster[sample(1:nrow(data_cluster), 1000), ]
hc <- hclust(dist(sample_data), method = "ward.D2")
fviz_dend(hc, k = 4, rect = TRUE)
km <- kmeans(data_cluster, centers = 4, nstart = 25)
fviz_cluster(km, data = data_cluster)
```

Результат:

Кластеризація показала, що фільми можна поділити на чотири групи зі спільними ознаками. Деякі кластери частково перекриваються, але загалом структура в даних простежується чітко.

Для оцінки впливу жанру, тривалості та року на успішність фільмів було побудувано логістичну регресію. Значущими виявилися тривалість, а також жанри Action, Comedy, Drama та Short. Короткометражки й драми частіше отримують високі оцінки, тоді як бойовики – рідше [2].

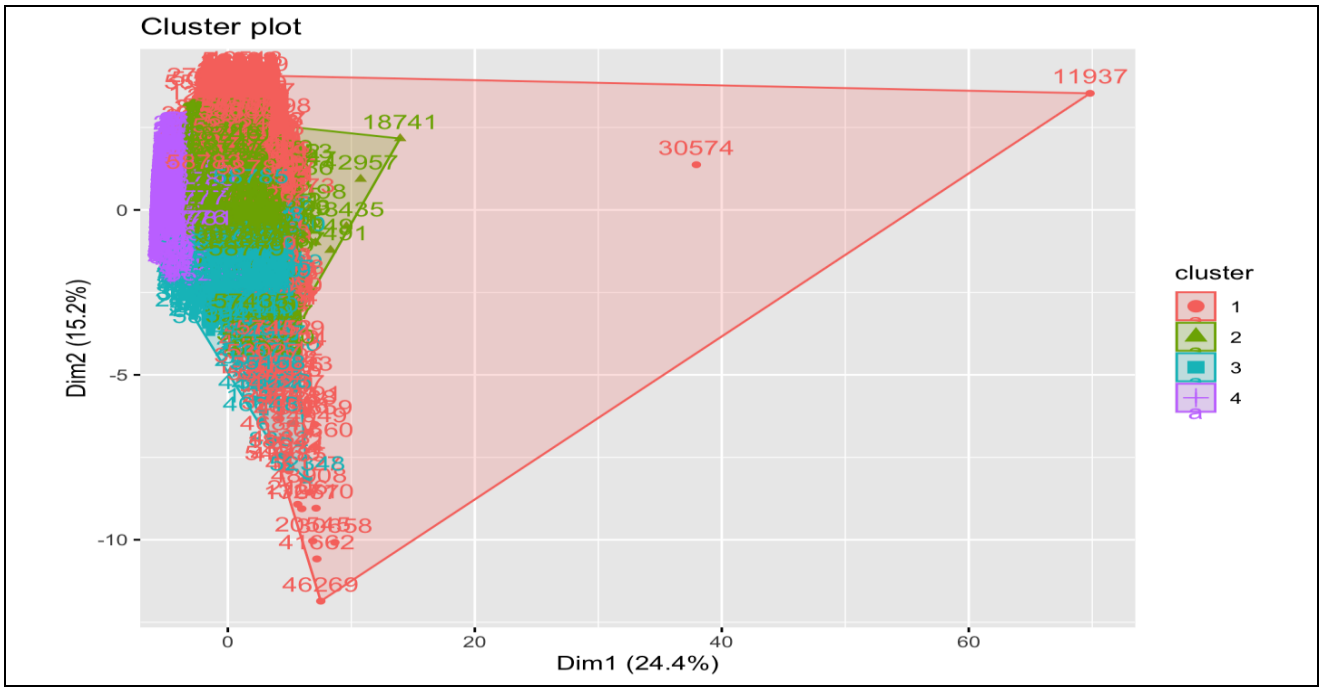

Рис. 2. Дендрограма кластерного аналізу

Лістинг програми:

```
movies_logit_data <- movies_clean %>%
  mutate(high_rating = rating > 7) %>%
  select(high_rating, year, length, Action, Comedy, Drama, Romance, Short) %>%
  na.omit()
logit_model <- glm(high_rating ~ ., data = movies_logit_data, family = binomial)
summary(logit_model)
```

Результат:

```
Call:
glm(formula = high_rating ~ ., family = binomial, data = movies_logit_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.9649357  0.8271136  -4.794 1.64e-06 ***
year         0.0007377  0.0004205   1.754   0.0794 .
length       0.0121181  0.0004772  25.393  < 2e-16 ***
Action      -0.7278759  0.0448627 -16.225  < 2e-16 ***
Comedy      -0.1236248  0.0231138  -5.349 8.87e-08 ***
Drama        0.2551355  0.0221252  11.531  < 2e-16 ***
Romance      0.0529373  0.0366527   1.444   0.1487
Short        1.8300761  0.0456593  40.081  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64618  on 58787  degrees of freedom
Residual deviance: 62283  on 58780  degrees of freedom
AIC: 62299

Number of Fisher Scoring iterations: 5
```

Рис. 3. Логістична регресія

Модель показала, що найбільше на ймовірність високої оцінки впливають тривалість, жанри Action, Comedy, Drama та формат Short. Короткометражки й драми частіше отримують високий рейтинг, а бойовики – рідше. Жанр Romance і рік випуску не мають значущого впливу.

Загалом кластеризація виявила чотири групи фільмів із подібними характеристиками, а логістична регресія підтвердила, що навіть базові ознаки можуть бути інформативними для оцінки успішності.

Список використаних джерел

1. Datasets. URL: https://www.kaggle.com/datasets
2. Освоюємо мову R. *Вікіпідручник.* 04.12.2025. URL: https://uk.wikibooks.org/wiki/%D0%9E%D1%81%D0%B2%D0%BE%D1%8E%D1%94%D0%BC%D0%BE_R
3. Кофанов О. Є., Солнцев С. О., Зозульов О. В. Програмування із використанням R у статистичних та маркетингових дослідженнях: навчально-методичний комплекс дисципліни: навч. посіб. для студентів спеціальності 075 «Маркетинг» / КПІ ім. Ігоря Сікорського. Київ: КПІ ім. Ігоря Сікорського, 2023. 204 с. URL: https://ela.kpi.ua/server/api/core/bitstreams/f78aa74c-7d8d-4c84-87eb-18a4aec57476/content
4. Data Science, лекція – Мова програмування R. *YouTube.* 03.05.2022. URL: https://www.youtube.com/watch?v=DiKwo_hgFfM

УДК 519.852:334.722

Менделюк К. В., *здобувачка вищої освіти 3 курсу спеціальності 122 Комп'ютерні науки,*
Хмелівський Ю. С., *асистент кафедри інформаційних технологій*

ЗАСТОСУВАННЯ МЕТОДІВ ОПТИМІЗАЦІЇ ДЛЯ ПЛАНУВАННЯ ФІНАНСОВИХ ІНВЕСТИЦІЙ

Донецький національний університет імені Василя Стуса, м. Вінниця

У сучасних умовах економічної нестабільності та високої конкуренції ефективне управління фінансовими ресурсами є одним із ключових завдань як для великих корпорацій, так і для дрібних інвесторів. Прийняття обґрунтованих інвестиційних рішень вимагає врахування великої кількості змінних: дохідності активів, рівня ризику, обмеженості ресурсів, ринкових коливань та макроекономічних факторів.

Методи оптимізації дають змогу формалізувати задачу вибору найкращого інвестиційного рішення, враховуючи всі важливі показники та обмеження. Завдяки цим методам можна розрахувати оптимальні варіанти розподілу ресурсів, підвищити дохідність інвестиційного портфеля та мінімізувати ризики.

Оптимізація у фінансових інвестиціях передбачає:

– формалізацію задачі у вигляді математичної моделі;

– визначення цільової функції (максимізація доходу, мінімізація ризику);

– врахування системи обмежень (обсяги капіталу, нормативи, ринкові фактори);