МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ СЕМЕНА КУЗНЕЦЯ

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КАФЕДРА ЕКОНОМІЧНОЇ КІБЕРНЕТИКИ І СИСТЕМНОГО АНАЛІЗУ

Рівень вищої освіти	Перший (бакалаврський)
Спеціальність	Системний аналіз
Освітня програма	Управління складними системами
Група	6.04.124.020.19.1

ДИПЛОМНИЙ ПРОЄКТ

на тему: «Прогнозування результатів спортивних ігор на основі методів багатовимірного аналізу»

Виконав: студент Максим СОЛОВЙОВ

Керівник: к.е.н., доцент Світлана ПРОКОПОВИЧ

Рецензент: к.т.н., доцент, директор
ТОВ «Юкрейніан Текнолоджи Трансфер Тим»
Ігор СОСНОВ

Харків – 2023 рік

РЕФЕРАТ

Звіт про дипломну роботу: 92 сторінки, 3 розділи, 52 малюнки, 1 таблиця, 47 джерел.

Об'єктом дослідження є процеси спортивної аналітики.

Мета дослідження – розробка моделі класифікації за допомогою методів багатовимірного аналізу, що дозволяє спрогнозувати результат матчу між двома обраними командами (результуюча змінна), а також розробка когнітивної моделі для аналізу факторів, що впливають на якість гри команд і запропонування дій, щодо поліпшення цієї якості.

Розглянуто концепцію та стратегію методів багатовимірного аналізу.

Результати дослідження зможуть дозволити прогнозувати результати футбольних матчів, що в подальшому може бути використано у різноманітних веб-сервісах зі спортивної аналітики.

КЛЮЧОВІ СЛОВА: СПОРТИВНА АНАЛІТИКА, ПРОГНОЗ, КЛАСТЕРНИЙ АНАЛІЗ, ДИСКРИМІНАНТНИЙ АНАЛІЗ, КОГНІТИВНА МОДЕЛЬ, СТАТИСТИКА, РІВЕНЬ ЯКОСТІ ГРИ.

ABSTRACT

Thesis of bachelor degree: 92 pages, 3 sections, 52 figures, 1 tables, 47 sources.

The object of research is the processes of sports analytics.

The purpose of the study is to develop a classification model using multivariate analysis methods that allows predicting the outcome of a match between two selected teams (resulting variable), as well as developing a cognitive model for analyzing factors affecting the quality of team play and proposing actions to improve this quality.

The concept and strategy of multivariate analysis methods are considered.

The results of the research will be able to predict the results of football matches, which can later be used in various web services for sports analytics.

KEY WORDS: SPORTS ANALYTICS, FORECAST, CLUSTER ANALYSIS, DISCRIMINANT ANALYSIS, COGNITIVE MODEL, STATISTICS, GAME QUALITY LEVEL.

ЗМІСТ

ВСТУП

З кожним днем стає дедалі більше вболівальників футболу. Цей вид спорту почали розвивати навіть у країнах Африки, Азії та Америки. У багатьох фанатів виникає не просто бажання подивитися на гру, а також додати азарту і заробити грошей. Але навіть найдосвідченіші букмекери не в змозі передбачати всіх випадковостей, які можуть статися до гри або під час гри. У міру розвитку технологій прогнозування футболу став очевидним той факт, що забезпечити стабільний виграш на ставках може лише суха статистика матчів. Як показує досвід успішних букмекерів, математичне моделювання набагато ефективніше, ніж будь-яка інша беттингова тактика.

Свого часу прогнозуваннями спортивних змагань та тематикою, близькою до цієї, займалися такі дослідники, як Капустін О., Штобва С. та інші [1-3]. Так, Капустін О. у своїх дослідженнях використав нейромережне прогнозування результатів тенісних матчів на основі аналізу природної мови. Штовба С. запропонував модель прогнозування футбольних матчів на основі теорії нечіткої логіки. Незважаючи на значний інтерес до досліджуваної тематики, питання порівняльної ефективності різних методів для прогнозування результату футбольного матчу, розглянуті недостатньо повно.

Метою роботи є розробка моделі класифікації за допомогою методів багатовимірного аналізу [4], що дозволяє спрогнозувати результат матчу між двома обраними командами (результуюча змінна), а також розробка когнітивної моделі для аналізу факторів, що впливають на якість гри команд і запропонування дій, щодо поліпшення цієї якості.

Для цього були поставлені та вирішені такі задачі:

розглянути поняття та особливості спортивної аналітики та прогнозування спортивних подій;

проаналізувати існуючі методи та моделі прогнозування результату футбольного матчу;

сформувати постановку задачі;

ознайомитися з суттю та завданням кластерного та дискримінантного аналізів, їх алгоритмами реалізації;

розробити модель прогнозування результату матчу за допомоги мови програмування R та середовища ППП Statistica;

запропонувати результат матчу і порівняти його з реальним рахунком;

розробити когнітивну модель аналізу критеріїв, що впливають на якість гри команд;

зробити висновки.

Об'єктом дослідження виступають процеси спортивної аналітики.

Предметом дослідження є методи та моделі кластерного та дискримінантного аналізу, а також моделі, побудовані на основі когнітивного моделювання.

Одним із найбільш популярних напрямів аналітики даних у сучасних умовах є спортивна аналітика. Спортивна аналітика цікава як для компаній-організаторів різних змагань з погляду параметризації формату заходів та рекламних кампаній, так і для учасників та глядачів, які нерідко роблять прогноз результатів спортивних подій. Тому дана тема, де досліджуються методи прогнозування результатів спортивних подій і способи аналізу та покращення якості гри команд, безумовно, є актуальною.

Результати дослідження зможуть дозволити прогнозувати результати футбольних матчів, що в подальшому може бути використано у різноманітних веб-сервісах зі спортивної аналітики.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ СПОРТИВНОЇ АНАЛІТИКИ

1.1. Поняття та особливості спортивної аналітики та прогнозування спортивних подій

Більшість уболівальників футболу, хокею, баскетболу та інших видів спорту дивляться матчі, зберігаючи всю інтригу і просто насолоджуючись видовищем. Але є і ті, для кого будь-яка спортивна подія – це низка статистичних даних, цифр, припущень і прогнозів. Це аналітики, які розглядають матчі як набір можливих подій, які можуть або не можуть відбутися [5].

Щоб зробити правильний прогноз, необхідно проаналізувати величезну кількість інформації та зробити висновки. Як правило, аналітики повинні вміти дивитися матчі та аналізувати статистику, вміти прогнозувати результати, розбирати та описувати події, робити аналіз статистичних даних тощо.

Якщо хтось стверджує про те, що його прогнози вірні на 100%, то перед вами або недосвідчений аналітик, або шахрай. Тим більше, якщо такі прогнози вам намагаються продати. Спортивні прогнози – це складна наука, яка не дає стовідсоткових припущень, тому ризикувати, покладаючись на прогнози, не можна [5].

Спортивна аналітика – це досить складний процес, який вимагає від аналітика як мінімум базових навичок для аналізу статистики, а також знання та розуміння спортивних подій. Адже тільки на основі даних можна робити досить точні прогнози. В даному випадку потрібно аналізувати спортивні новини, які допоможуть досягти успіху.

Для повного розуміння спортивної аналітики потрібно також розглянути поняття прогнозування. Прогнозування [6] (грец. prognosis – знання наперед) – це процес формування прогнозів на основі аналізу тенденцій і закономірності розвитку об'єкта.

Необхідність прогнозу обумовлена бажанням знати події майбутнього, що достовірно – неможливо в принципі, виходячи із статистичних (помилки

поточних оцінок), ймовірнісних (багатоваріантність наслідків), емпіричних (методологічні помилки моделей), філософських (обмеженість поточних знань) принципів. Щоб уміти аналізувати ту чи іншу подію, потрібно бути справжнім гуру в цьому виді спорту, знати явні і неявні фактори, мати певний склад розуму і талант. Але навіть в цьому випадку завжди буде щось непередбачене, що може зробити прогноз в корені помилковим [5].

Прогнозування зобов'язане, таким чином, відповісти на два питання: чого ймовірніше всього можна очікувати в майбутньому і яким чином потрібно впливати на умови, щоб досягнути заданої мети (стану) [28].

Прогностика - це наука, яка вивчає закономірності процесу прогнозування. Предметом прогностики є дослідження законів і способів прогнозування [28].

Прогноз [7] – це ймовірнісне судження про майбутній стан об'єкта дослідження – результат матчу.

Точність будь-якого прогнозу обумовлена:

1. Доступністю та якістю вихідних даних: точність прогнозу значно залежить від доступності і якості вихідних даних. Якщо вхідні дані неповні, неточні або спотворені, прогноз може бути неточним або недостовірним.

2. Моделлю або методологією прогнозування: використання правильної моделі або методології прогнозування є важливим фактором для досягнення точних результатів. Різні моделі мають свої обмеження.

3. Урахування змінних та факторів: точність прогнозу може залежати від того, наскільки добре модель або методологія враховує різні змінні та фактори, які можуть впливати на майбутні події.

4. Якістю аналізу та обробки даних: аналіз та обробка даних перед прогнозуванням може суттєво вплинути на його точність. Якщо дані не аналізуються належним чином, помилки можуть виникати на етапі підготовки даних і призводити до неточності прогнозу.

Отже, на основі всіє зібраної інформація, ми маємо змогу побудувати блок-схему основних операцій розробки прогнозної моделі у футбольній сфері, де об’єктом є дві обрані команди та результат гри між ними (рис. 1.1).

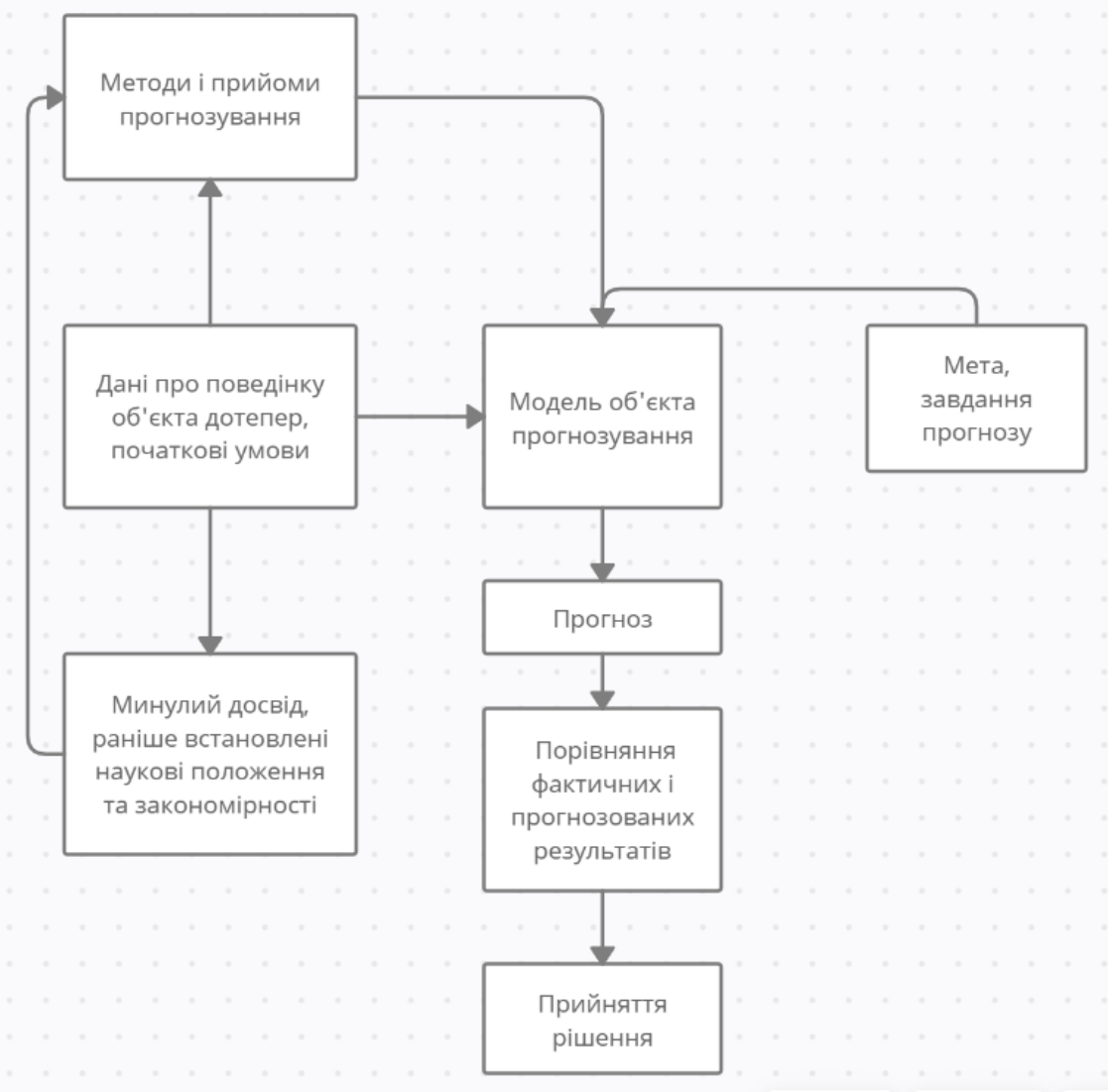

Рис. 1.1. Основні операції розробки прогнозної моделі у футболі

Таким чином, ми можемо переглянути всі основні операції під час розробки прогнозної моделі у футбольній сфері на побудованій блок-схемі.

1.2. Огляд існуючих методів та моделей прогнозування результату футбольного матчу

Дослідження в галузі прогнозування футбольних результатів виконувалися такими вченими, як С. Добсон, А. Ротштейн, Дж. Годдард та С. Штовба. Стівен Добсон та Джон Годдард розробили модель, в якій головним фактором було

кількість голів, забитих кожною командою у особистих зустрічах. Ця модель була побудована на основі даних з близько 30 сезонів [8].

Інший вчений, А. Ротштейн, використовував алгоритм ставок на основі нечіткої логіки, використовуючи дані з 12 чемпіонатів Фінляндії. Розрахунок ставок на футбол виконувався за допомогою нейронних мереж та генетичного навчання [8].

С. Штовба прогнозував розподіл місць у турнірній таблиці українського футбольного чемпіонату на основі нечіткої логіки [8].

Буурсма вибрав набір функцій і використав різні алгоритми класифікації, включаючи просту і логістичну регресію, байєсівську мережу та дерево прийняття рішень, для передбачення результату футбольних матчів. Його прогнозування мали три можливих варіанти (перемога домашньої команди, нічия, перемога гостей), і ймовірності цих трьох результатів були розраховані, обираючи результат з найбільшою імовірністю [9].

Протягом багатьох років розроблялися моделі та комп'ютерні програми для передбачення результатів спортивних ігор. Більшість з них використовує стохастичні методи для врахування невизначеності, такі як регресивний і авторегресивний аналіз, метод Байезіана в поєднанні з ланцюгами Маркова і метод Монте-Карло. Особливості таких моделей полягають у їх високій складності, потребі у великій кількості статистичних даних та наявності численних припущень. Крім того, ці моделі не завжди легко інтерпретуються [7].

Крім стохастичних підходів, також існують моделі, які використовують нейронні мережі для передбачення результатів футбольних матчів. Ці моделі засновані на аналізі великого обсягу даних та використанні штучних нейронних мереж для виявлення складних залежностей та патернів у цих даних. Підхід з використанням нейронних мереж може мати свої переваги щодо точності передбачень і здатності адаптуватись до нових ситуацій. [7].

Для зручного порівняння методів прогнозування футбольних матчів сформулюємо табл. 1.1, в якій наведемо перелік усіх вищеописаних методів та моделей з результатами, які досягли дослідники.

Таблиця 1.1

Аналіз методів прогнозування футбольних подій

№	Метод	Результат моделювання
1	2	3
1	Метод прогнозування футбольних змагань на основі статистичного аналізу та кваліметрії [10]	Було розроблено комп'ютерну систему, яка демонструє великий потенціал у прогнозуванні результатів футбольних матчів. Шляхом проведення серії експериментів, використовуючи цей метод, було досягнуто високої точності прогнозування, наближеної до 90%. Аналізуючи результати цих експериментів, було встановлено, що основи кваліметрії та статистичного аналізу є найефективнішими для вирішення завдання прогнозування результатів футбольних матчів. Ця комп'ютерна система заснована на глибокому аналізі великої кількості статистичних даних, включаючи результати попередніх матчів, статистику команд та гравців, фактори впливу, такі як погодні умови і стан поля, а також інші релевантні параметри. Завдяки цій системі тренери, гравці та футбольні експерти мають можливість отримувати обґрунтовані прогнози, які допомагають у прийнятті рішень, формуванні стратегій та підготовці до матчів. Такий підхід сприяє покращенню якості аналізу та підвищує шанси на досягнення успіху в футболі [13].
2	Прогнозування результатів футбольних матчів за допомогою методу опорних векторів (SVM) [8]	На сьогоднішній день відсутні достатньо значущі наукові дослідження, що б дозволили побудувати одноітераційну модель опорних векторів (SVM) для прийняття рішень в багатьох класах. У зв'язку з цим, вихідна задача прогнозування результатів футбольних матчів з трьома можливими класами рішень (перемога команди-господаря поля, нічия та перемога гостьової команди) була змінена на типову регресійну задачу, в якій використовувався SVM-алгоритм для мінімізації середньоквадратичної помилки. Перехід від неперервного значення вихідного показника моделі до дискретного результуючого значення здійснюється за такими правилами: якщо прогнозоване значення не є від'ємним, то результат матчу визначається як "гостьова команда не переможе"; якщо прогнозоване значення є від'ємним, то результат матчу визначається як "команда-господар поля не переможе". Іншими словами, прогноз результату футбольного матчу здійснюється на основі знаку вихідного показника. Цей підхід дозволяє зменшити складність задачі і забезпечити більш просту та ефективну модель для прогнозування результатів футбольних матчів [13].
3	Метод прогнозування за допомогою теорії нечітких множин [11]	Була проведена відносно невелика вибірка, в яку входило всього 175 футбольних матчів, з метою зробити висновки щодо прогнозування результатів. Система продемонструвала високий рівень точності прогнозування - 64%, при врахуванні трьох можливих

Закінчення табл. 1.1

1	2	3
		результатів (перемога, поразка, нічия).Проте, систему не було протестовано з використанням коефіцієнтів. Для належного функціонування системи необхідне налаштування коефіцієнтів, які використовуються для перерахунку термів. Це означає, що у випадку використання інших даних або великого обсягу матчів, необхідно врахувати та налаштувати коефіцієнти, щоб забезпечити точність та надійність прогнозів. Тестування системи з різними наборами коефіцієнтів допоможе встановити оптимальні значення та забезпечити її ефективну роботу [13].
4	Метод зваженої суми показників для прогнозування футбольних подій [12]	Розроблена математична модель може бути використана для прогнозування результатів футбольних матчів. Аналітик повинен провести детальний аналіз статистичних показників команд та врахувати свій власний досвід, щоб визначити вагові коефіцієнти і, можливо, виключити деякі показники з розрахунку. Це означає, що перед прогнозуванням кожного конкретного матчу аналітик повинен внести необхідні корективи до моделі, враховуючи важливість певних показників та свої експертні знання. Наприклад, він може приділити більше уваги показникам, які мають вирішальне значення, таким як форма гравців, травмованість, історія протистоянь команд тощо. Такий підхід дозволяє покращити точність прогнозування, оскільки враховує індивідуальні особливості кожного матчу та контекст, в якому він відбувається. Аналітик, як експерт у галузі, може додати свої інтуїтивні оцінки та знання, що сприяє більш точному та обґрунтованому прогнозуванню результатів футбольних матчів [13].

Таким чином, були розглянуті і проаналізовані різноманітні методи прогнозування футбольних подій.

1.3. Постановка задачі

В межах дипломного проєкту запропоновано розробити моделі класифікації, використовуючи методи багатовимірного аналізу, з метою прогнозування результату матчу між двома вибраними командами. Крім того, пропонується створити когнітивну модель для аналізу факторів, що впливають

на якість гри команд, та запропонувати шляхи покращення цієї якості [29].

Моделі класифікації, розроблені за допомогою методів багатовимірного аналізу, дозволять систематично оцінювати параметри, що впливають на результат матчу. Вони будуть засновані на комплексному аналізі різних факторів, таких як статистика команд, форма гравців, тактика гри та інші. Це дозволить зробити більш точний прогноз та допоможе в розумінні динаміки футбольних поєдинків [29].

Крім того, когнітивна модель буде розроблена з метою аналізу та оцінки чинників, що впливають на якість гри команд. Це можуть бути такі фактори, як координація між гравцями, стратегія команди, підготовка футболістів та багато інших. Аналіз цих чинників допоможе ідентифікувати сильні та слабкі сторони команди та розробити рекомендації щодо покращення їхньої гри [29].

Загальна мета проєкту полягатиме в розробці інструменту, який допоможе тренерам та фахівцям у футбольній сфері приймати обґрунтовані рішення та покращувати якість гри команд. Це сприятиме досягненню більш високих результатів та розширенню знань у галузі футбольного аналізу [29].

Для реалізації моделей спершу потрібно провести детальний аналіз футбольної галузі. Обрати чемпіонат та окремий матч, на який будемо робити прогноз (такий, який вже був зіграний, для того, щоб була можливість порівняти прогнозоване значення з реальним і зробити висновки). Так, як будь-який матч несе в собі безліч різної інформації, виокремимо такі найважливіші показники [14]:

інформація про матч в цілому (які команди грають, їх склад, рахунок, дата проведення, стадіон);

статистика за гру (забиті – пропущені голи, отримані картки, кількість часу, проведеного в грі, кількість ударів, кутових, передач тощо, відсоток контролю м'яча);

На рис. 1.2–1.3 можна побачити приклад опису конкретного матчу за усіма вище перерахованими критеріями.

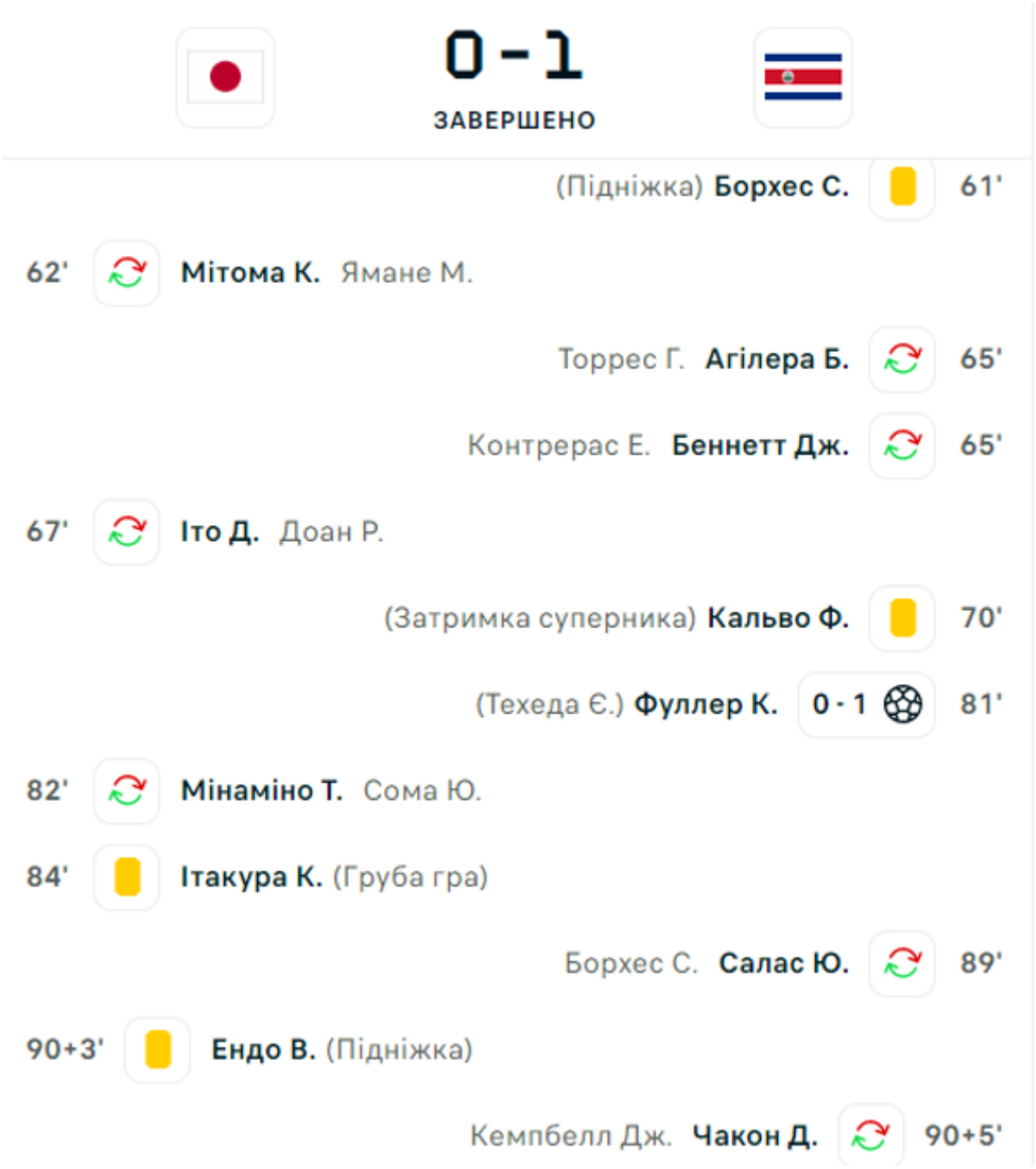

Рис. 1.2. Хронологія матчу

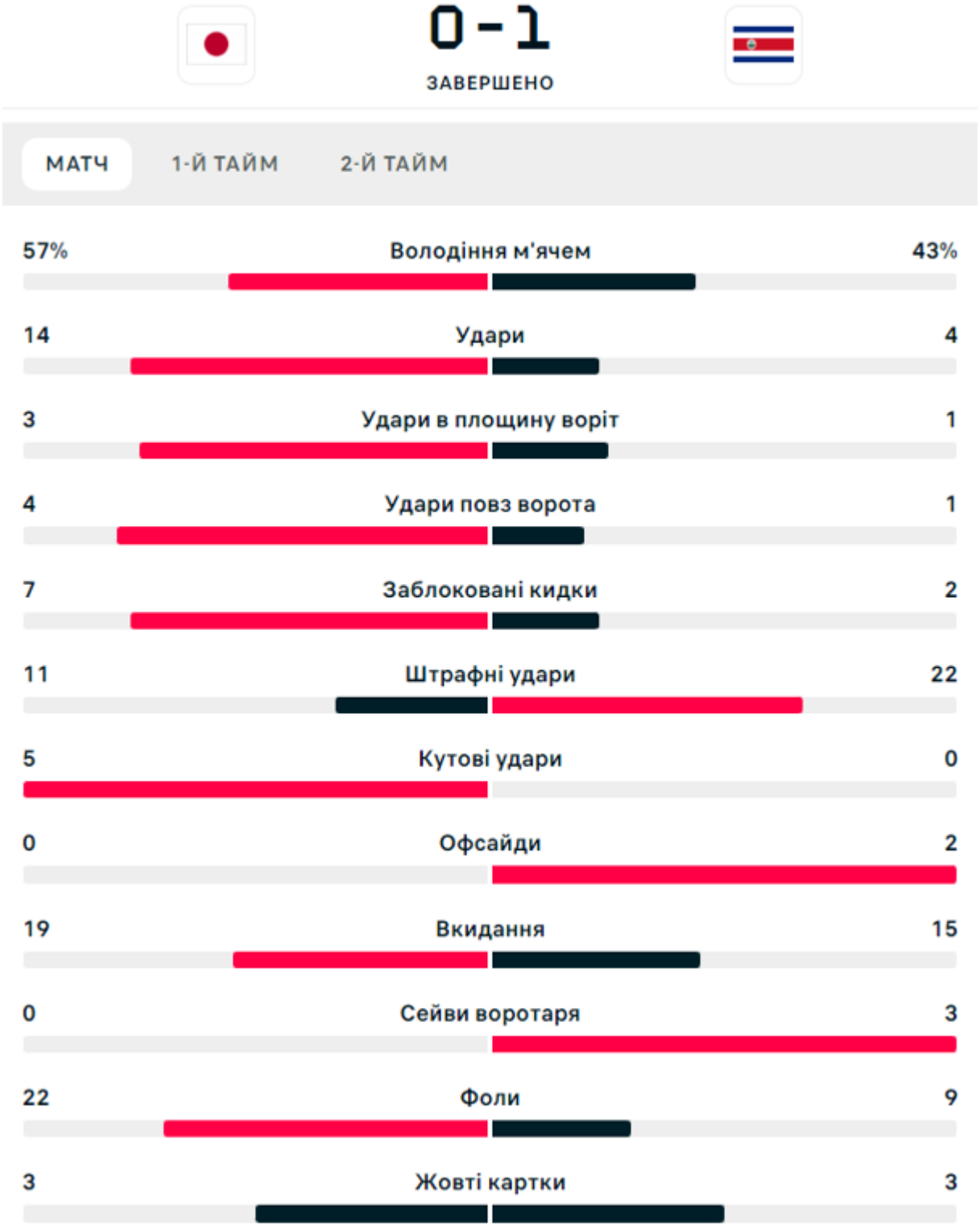

Рис. 1.3. Статистика за матч

Однак більш корисною і популярною для футбольних фанатів є інформація не за окремим матчем або командою, а таблиця результатів певного чемпіонату – рис. 1.4, яку зазвичай, представляють у вигляді рейтингового списку досягнень команд.

# ▲	КОМАНДА	І	В	Н	П	Г	ОЧ.	ФОРМА
1.	Дніпро-1	14	11	2	1	31:10	35	? П ? В Н В
2.	Шахтар Д.	13	9	3	1	29:11	30	? В П Н В В
3.	Зоря	14	8	3	3	30:20	27	? Н В Н П В
4.	Динамо К.	14	8	3	3	21:13	27	? Н В В Н В
5.	Олександрія	13	7	3	3	23:19	24	? Н ? П В П
6.	Колос	14	5	4	5	12:16	19	? В П Н П П
7.	Металіст 1925	14	4	6	4	13:17	18	? Н В В Н П
8.	Ворскла	15	5	3	7	18:20	18	? П В Н П В
9.	Верес	15	5	2	8	17:20	17	? П П П П В
10.	Металіст	14	4	4	6	15:20	16	? Н П П Н П
11.	Кривбас	14	4	3	7	10:16	15	? В П В В П
12.	Минай	14	3	5	6	10:15	14	? Н В Н П П
13.	Рух	13	3	4	6	15:20	13	? Н Н П В Н
14.	Інгулець	14	3	4	7	12:18	13	? В В П П В
15.	Чорноморець	13	2	4	7	8:17	10	? Н В В П П
16.	Львів	14	2	3	9	9:21	9	? П П П В П

Рис. 1.4. Таблиця «Чемпіонату України 2022-2023»

Таблиця відображає кількість зіграних ігор (І), яка на кінець туру, зазвичай, повинна бути однаковою для всіх команд. Також матчі, які були виграні (В), зіграні в нічию (Н) і програні (П). Окрема колонка (Г) виділена під співвідношення між забитими і пропущеними голами.

Записи в таблиці упорядковано відповідно до зменшенням кількості набраних командою очок, які розраховуються за формулою:

$$О = В \cdot 3 + Н \cdot 1 + П \cdot 0,$$

де В – кількість перемог;

Н – кількість нічиїх;

П – кількість поразок.

Колонка «Форма» показує те, як команда зіграла минулі 5 матчів.

Для нашого аналізу і побудови моделі знадобляться дані як з загальної таблиці, так і дані за окремо взятою грою.

Наступним кроком є пошук інформації, та способу реалізації спочатку кластерного, а потім дискримінантного аналізу. На даному етапі можна визначити наступні пункти:

розробити математичне представлення задачі прогнозування футбольних матчів;

проаналізувати метод кластерного аналізу;

дослідити метод дискримінантного аналізу;

сформулювати математичну постановку задачі для прогнозування результатів футбольних матчів за допомогою даних методів.

Далі постає власне етап розробки моделі прогнозування футбольного матчу. Тобто реалізація статистичного моделювання [30], аналіз моделі прогнозування, порівняння отриманих результатів з реально існуючими і формулювання висновків за поточними моделями.

Наступним, і завершальним завданням, постає реалізація когнітивної моделі, за допомогою якої буде можливе проведення аналізу статистичних критеріїв, що можуть впливати на якість гри команд, та, звичайно, формулювання висновків щодо подальшої можливості покращення цієї якості.

РОЗДІЛ 2. МЕТОДИ ТА МОДЕЛІ БАГАТОВИМІРНОГО АНАЛІЗУ І КОГНІТИВНОГО МОДЕЛЮВАННЯ У ОЦІНЮВАННІ ЯКОСТІ ГРИ КОМАНД ТА ПРОГНОЗУВАННІ РЕЗУЛЬТАТУ ФУТБОЛЬНОГО МАТЧУ

2.1. Особливості класифікації багатовимірних об’єктів

У дослідженнях, кількість об'єктів аналізу може бути досить великою - від десятків до сотень, і кожен об'єкт характеризується десятками ознак. Очевидно, що прямий (візуальний) аналіз такої великої матриці даних стає малоефективним. Тому виникає потреба у зведенні даних, концентрації і аналізі структури об'єктів дослідження. Сучасні методи багатовимірної класифікації можуть бути використані для вирішення цих завдань [19].

Методи багатовимірної класифікації дозволяють групувати об'єкти з урахуванням всіх суттєвих структурно-типологічних ознак та розподілу об'єктів у заданій системі ознак. Метою такої класифікації є об'єднання схожих об'єктів у одну групу, з максимальною відмінністю між об'єктами з різних груп [19].

Таким чином, методи багатовимірної класифікації використовуються для розділення сукупності об'єктів на однорідні групи, при цьому кожен об'єкт характеризується значним числом стохастично пов'язаних ознак [19].

Під час досліджень зазвичай використовують певні підходи до класифікації об'єктів, а також відповідні методи класифікації, які наведені на рис. 2.1.

Розглянемо основні терміни кластерного аналізу:

Кластер - це група об'єктів, які мають подібні характеристики або властивості [19].

Кластерний аналіз (або аналіз кластерів) - це статистичний метод, що використовується для групування схожих об'єктів разом в кластери. Його основна мета - виявити приховані структури або групи в наборі даних, де об'єкти всередині кожної групи подібні один до одного, а об'єкти між групами відрізняються [19].

Кластерний аналіз зазвичай використовується для розуміння внутрішньої структури даних, знаходження груп подібних об'єктів та виявлення закономірностей. Він може бути застосований до різних областей, включаючи науку про дані, маркетингове дослідження, медичну діагностику, соціологію, географію та багато інших [19].

У процесі кластерного аналізу використовуються різні методи і алгоритми, такі як метод k-середніх, ієрархічна кластеризація, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) та інші. В результаті аналізу отримується групування об'єктів в кластери, що допомагає зрозуміти схожість та різноманітність об'єктів у наборі даних і дозволяє виявити внутрішні зв'язки та структури [19].

Кластерний метод (або метод кластеризації) є алгоритмом аналізу даних, який використовується для групування схожих об'єктів в кластери на основі їх схожості або відстані один від одного. Кластерний метод може бути застосований до різних видів даних, таких як числові дані, категоріальні дані або текстові дані, з метою виявлення прихованих структур або груп в наборі даних. Цей метод є популярним інструментом для візуалізації та розуміння взаємозв'язків в даних, а також для виявлення нових знань та патернів [19].

Рис. 2.1. Підходи до класифікації об’єктів

Кластерний аналіз використовується у ході проведення економічних досліджень з метою вирішення різних завдань, що ґрунтуються на багатовимірній класифікації. Основні задачі, які можуть бути розв'язані за допомогою кластерного аналізу можемо побачити на рис.2.2 [19].

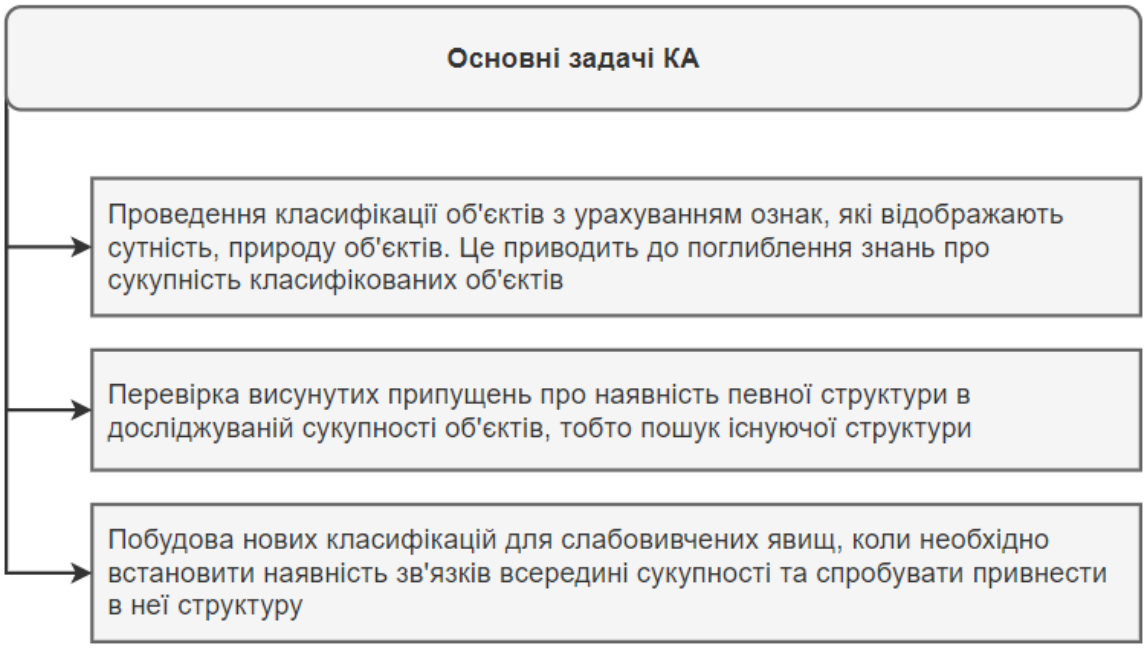

Рис. 2.2. Основні задачі кластерного аналізу

Кластерний аналіз застосовується у багатьох галузях і областях, де виникає необхідність у групуванні об'єктів за схожістю або виявленні прихованих структур у наборі даних. Основні області застосування кластерного аналізу включають [19]:

Маркетингові дослідження: Кластерний аналіз використовується для сегментації ринку, ідентифікації груп споживачів зі схожими характеристиками та визначення специфічних маркетингових стратегій для кожної групи [19].

Соціологічні дослідження: Кластерний аналіз допомагає ідентифікувати субгрупи населення з подібними соціальними, економічними або демографічними характеристиками. Це може бути корисно для визначення цільових аудиторій, розробки соціальних програм або вивчення політичних настроїв [19].

Медичні дослідження: Кластерний аналіз застосовується для класифікації пацієнтів на основі медичних даних та виявлення схожих патологічних зразків. Це може допомогти у прогнозуванні ризиків, виборі оптимального лікування та розробці індивідуальних підходів до кожного пацієнта [19].

Фінансовий аналіз: Кластерний аналіз допомагає ідентифікувати схожі групи компаній або фінансових інструментів на основі фінансових показників. Це може бути корисно для портфельного управління, ризик-аналізу та прийняття інвестиційних рішень [19].

Географічний аналіз: Кластерний аналіз використовується для групування географічних об'єктів, таких як населені пункти або регіони, залежно від їхньої географічної близькості або подібних характеристик. Це може бути корисно для регіонального планування, маркетингу на місцевому рівні та аналізу розподілу ресурсів [19].

Біологічні дослідження: Кластерний аналіз використовується для класифікації організмів на основі генетичних або молекулярних характеристик. Це може допомогти у вивченні еволюційних зв'язків, виявленні нових видів або ідентифікації генетичних дефектів [19].

У нашому ж випадку, кластерний аналіз використовується для розбиття на групи футбольних команд за статистичними характеристиками.

Кластерний аналіз визначає групи об'єктів, які знаходяться далеко один від одного та компактно розташовані, і шукає "природне" поділ сукупності на області, де об'єкти сконцентровані. Цей метод використовується, коли вихідні дані представлені у вигляді матриць схожості або відстаней між об'єктами, або як точки у багатовимірному просторі. Найпоширенішим типом даних є останній, для яких кластерний аналіз спрямований на виявлення геометрично віддалених груп, в межах яких об'єкти є близькими один до одного [20].

Внаслідок використання методів кластерного аналізу, об'єкти дослідження групуються в окремі кластери, які проявляють такі характеристики (рис. 2.3) [20]:

Рис. 2.3. Властивості кластера

Розглянемо типи кластерних структур.

Визначення терміну "типу кластерної структури" не має чіткого формального опису і залежить від нормування ознак та використаного методу кластеризації. Нижче наведено приклади різних типів кластерних структур, які можуть виникати при проведенні аналізу (кластери з центром, стрічкові кластери, кластери з перемичками, перекриваючі кластери, кластери, що виникають на основі різних типів регулярності, а також відсутність кластерів), як показано на рис. 2.4 [20]:

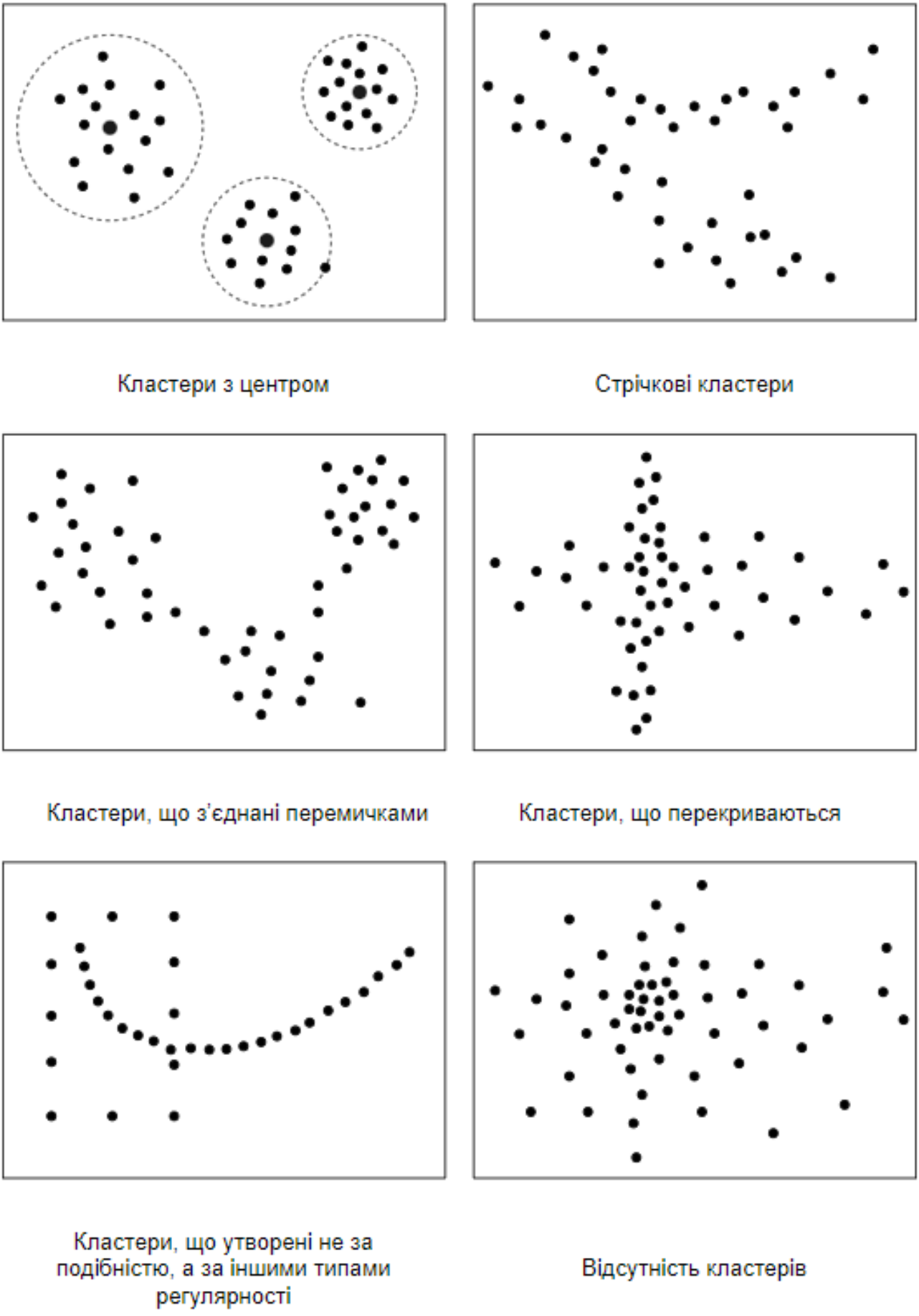

Рис. 2.4. Типи кластерних структур

Також, розглянемо формальну постановку задачі кластеризації. Нехай X – множина об'єктів, Y – множина номерів (імен) кластерів. Задана функція відстані між об'єктами $\rho(x,x')$. Сформована навчальна вибірка об'єктів $X^m = x_1, \dots x_m \subset X$. Необхідно розбити вибірку на непересічні підмножини, які називають кластерами, так, щоб кожен кластер складався з об'єктів, близьких за метрикою ρ, а об'єкти різних кластерів істотно відрізнялися. Кожному об'єкту $x_i \in X^m$ приписується номер кластера Y_i [20].

У навчальній вибірці об'єкти можуть характеризуватися ознаками, які вимірюються в різних одиницях. Однак для кластерного аналізу ознаки повинні бути однорідними, тобто вимірюватися в порівняльних шкалах. Для цього здійснюється нормування початкових даних (рис. 2.5) [20]:

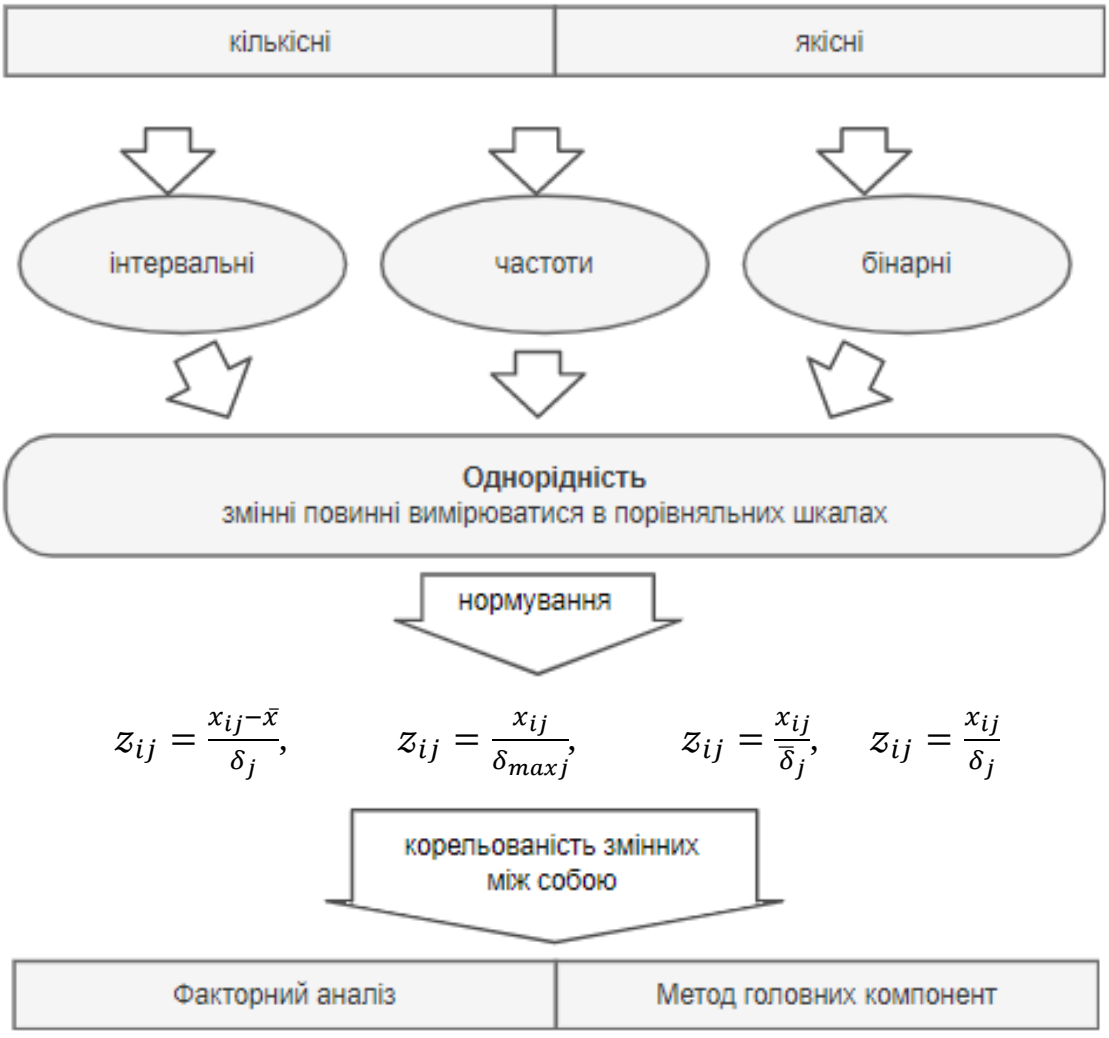

Рис. 2.5. Вимоги до початкових даних

Отже, можемо сформувати основні етапи кластерного аналізу [20]:

1. Підготовка даних: Обробка та підготовка вхідних даних, включаючи нормалізацію, видалення аномальних значень і обробку пропущених даних.

2. Вибір методу кластерного аналізу: Вибір підходящого методу аналізу, який враховує характер досліджуваних даних і мету аналізу.

3. Визначення кількості кластерів: Встановлення кількості кластерів, яку потрібно виділити у досліджуваних даних.

4. Виконання кластеризації: Застосування вибраного методу кластерного аналізу до підготовлених даних з метою виділення кластерів.

5. Валідація кластерів: Оцінка якості отриманих кластерів за допомогою різних метрик та методів валідації, таких як внутрішньокластерна і міжкластерна варіація, індекси схожості, статистичні тестування тощо [20].

Давайте розглянемо більш детально один з ключових етапів, а саме обчислення міри подібності між об'єктами. Цей етап є важливим у процесі кластеризації, оскільки ми хочемо, щоб об'єкти зі схожими характеристиками потрапляли в один кластер. Для досягнення цієї мети ми використовуємо різні міри подібності у кластерному аналізі (рис. 2.6) [21]:

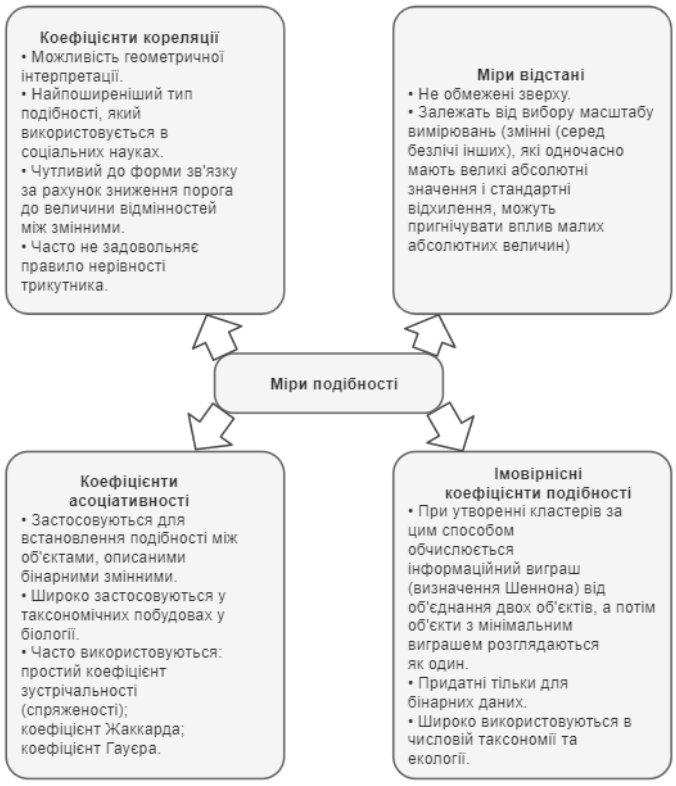

Рис. 2.6. Міри подібності в кластерному аналізі

У процесі кластеризації кожен об'єкт розглядається як точка у багатовимірному просторі ознак, які використовуються для опису цих об'єктів. Одним із ключових аспектів при класифікації є визначення подібності та відмінності між цими точками, а це досягається за допомогою метричних відстаней між ними [21].

Метричні відстані представляють собою числові значення, які вимірюють віддаленість або схожість між двома об'єктами в багатовимірному просторі. Ці відстані дозволяють визначити, наскільки близькими або віддаленими є об'єкти один від одного з точки зору їх ознак або характеристик [21].

Різні метричні відстані можуть використовуватися в залежності від типу даних та вимог задачі. Наприклад, Евклідова відстань вимірює прямолінійну відстань між двома точками, тоді як косинусна схожість враховує кут між векторами, що представляють об'єкти [21].

Використання відповідної метричної відстані є важливим кроком у процесі кластеризації, оскільки вона дозволяє визначити близькість чи віддаленість між об'єктами. Це допомагає згрупувати подібні об'єкти в одній кластері та виділити відмінності між різними кластерами [21]. Це дозволяє встановити ступінь схожості між об'єктами та виміряти відстань між ними на основі відповідних ознак:

- введенням правила обчислень відстаней $d(x_i,x_j)$ між будь-якою парою досліджуваних об’єктів (x_1, x_2, …, x_n)
- заданням деякої функції $r(x_i,x_j)$, що характеризує ступінь близькості i-го і j-го об’єктів [21].

Міра подібності є метрикою, якщо виконуються наступні умови [21]:

1. Симетрія. Відстань між об’єктами x і y повинна задовольняти:

$$d(x, y) = d(y, x) \geq 0;$$

2. Нерівність трикутника. Відстань між об’єктами x, y і z:

$$d(x,y) \le d(x,z) + d(y,z);$$

3. Розрізненість нетотожних об’єктів. Дано два об’єкта *x* і *y*, якщо:

$$d(x,y) \neq 0, \text{ то } x \neq y;$$

4. Нерозрізненість ідентичних об’єктів. Якщо *x* і *x′* ідентичні, то:

$$d(x,x') = 0.$$

Розглянемо міри подібності, які використовують у кластерному аналізі: коефіцієнт кореляції, ймовірнісний коефіцієнт подібності, міри відстані, коефіцієнти асоціативності [21].

1. Коефіцієнт кореляції – природна міра подібності [21]:

$$r_{ij} = \frac{\sum_{h=1}^{N}(x_{hi}-m_i)(x_{hj}-m_j)}{\delta_i \delta_j},$$

$$-1 \le r_{ij} \le 1,$$

де x_{hi}, x_{hj} – значення *h*-ї ознаки для *i*-го та *j*-го об’єктів;

$m_i, m_j, \delta_i, \delta_j$ – відповідні середні та середньоквадратичні відхилення для характеристик *i* та *j*.

При $r_{ij} = -1$ – наявність зворотного тісного зв’язку між об’єктами *i* і *j*;

$r_{ij} = 0$ – відсутність зв’язку між об’єктами *i* і *j;*

$r_{ij} = 1$ – наявність прямого тісного зв’язку між об’єктами *i* і *j* [21].

2. Імовірнісний коефіцієнт подібності – міра близькості імовірнісного типу [21]:

$$I_{ij} = \sum_{x,y} p_{xy} log \frac{p_{xy}}{p_x^i p_y^j},$$

де p_{xy} – ймовірність спільної появи ознак *x* і *y*;

p_x^i – ймовірність появи ознаки *x* в об'єкті *i*;

p_y^j – ймовірність появи ознаки *y* в об'єкті *j* [21].

3. Міри відстані – у кластерному аналізі використовують наступні міри відстані: евклідова відстань, «зважена» евклідова відстань, City-blok (Мангетенська), відстань Мінковського [21].

Евклідова:

$$d_{ij}\sqrt{\sum_{k=1}^{m}\left(x_{ik}-x_{jk}\right)^2},$$

де d_{ij} – відстань між об'єктами *i* і *j*, *i*, *j* = 1, …, *n*; *k*=1, …, *m*;

x_{ik} – значення *k*-ї змінної для *i*-го об'єкта;

x_{jk} – значення *k*-ї змінної для *j*-го об'єкта [21].

«Зважена» Евклідова:

$$d_{ij}=\sqrt{\left(\sum_{k=1}^{m}\omega_k\cdot\left(x_{ik}-x_{jk}\right)^2\right)},$$

де ω_k – вага k-ї ознаки, $0\le\omega\le 1$ [21].

City-blok (Мангетенська):

$$d_{ij}=\sum_{k=1}^{m}\left|x_{ik}-x_{jk}\right|.$$

Відстань Мінковського:

$$d_{ij}=\left(\sum_{k=1}^{m}\left|x_{ik}-x_{jk}\right|^p\right)^{1/r},$$

де p, r – параметри, що визначені користувачем [21].

Відстань Махаланобіса:

$$d_{ij} = \left(X_i - X_j\right)^T \cdot S^{-1} \cdot \left(X_i - X_j\right),$$

де X_i, X_j – вектори значення *i*-го та *j*-го об'єктів;

S – загальна коваріаційна матриця [21].

4. Коефіцієнти асоціативності – для бінарних даних, змінні, що беруть участь в конструюванні цих заходів, описуються таблицею асоціативності, де «1» вказує на наявність змінної, а «0» – її відсутність (рис. 2.7) [21]:

	1	0
1	a	b
0	c	d

Рис. 2.7. Змінні для конструювання коефіцієнтів асоціативності

Простий коефіцієнт зустрічності:

$$S = \frac{(a+d)}{(a+b+c+d)}.$$

Коефіцієнт Жаккара:

$$J = \frac{a}{(a+b+c)}.$$

Коефіцієнт Гауера:

$$S_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}},$$

де S_{ijk} – «вклад» у подібність об’єктів, який враховує значущість ознаки k, у порівнянні об’єктів i та j;

W_{ijk} – ваговий коефіцієнт, який приймає значення 1 – якщо порівняння об’єктів за ознакою k варто враховувати і 0 – в іншому випадку [21].

2.2. Сутність, завдання та алгоритм реалізації дискримінантного аналізу

Дискримінантний аналіз (discriminant analysis) – метод багатовимірного статистичного аналізу. Він включає методи класифікації багатовимірних спостережень за принципом максимальної подібності за наявності навчальних ознак. На відміну від кластерного аналізу, нові кластери не утворюються, а є правилом, за яким об'єкти відносяться до певної групи. Завдання дискримінантного аналізу багато в чому схожі на завдання логістичної регресії – класифікація спостережень на групи на основі прогностичної моделі. Незважаючи на деякі подібності, дискримінантний аналіз і логістична регресія мають істотні відмінності. Ідеї дискримінантного аналізу тісно пов'язані з дисперсійним, регресійним аналізом [15].

Сутність дискримінантного аналізу [16] – виходячи з навчальних вибірок перетворити багатовимірний масив на одновимірний показник для прогнозування належності спостережень до груп, тобто побудувати новий узагальнений показник, значення якого максимально різняться для об'єктів, віднесених до різних груп. Навчальна вибірка [16] – це безліч об'єктів, в нашому випадку команд, заданих значеннями статистичних показників та належність яких до того чи іншого класу достовірно відома.

Дослідження різниці між групами – основа концепції дискримінантного аналізу. У нашому прикладі, залежна змінна може бути результатом команди в матчі (В – виграш, Н – нічия або П – поразка), а незалежними змінними можуть

бути характеристики команд, які грають між собою. Під час проведення дискримінантного аналізу – знаходять дискримінантну функцію (лінійну комбінацію незалежних змінних), яка найкраще розрізняє категорії чи групи залежної змінної [16].

Дискримінантний аналіз проводиться в умовах наступних основних припущень [31]:

множина M об'єктів розбита на два або більше ($q >= 2$) підмножин M_k (класу), які відрізняються від інших груп змінними x_{ij};

у кожній підмножині M_k знаходиться, принаймні, два об'єкти ($n_k >= 2$), причому всі об'єкти спостереження множини M повинні належати якійсь із підмножин (класів);

число N об'єктів спостереження має перевищувати число p дискримінантних змінних ($0 < p < N - 2$) не менше ніж на дві одиниці;

лінійна незалежність між ознаками (j), тобто жодна з ознак не повинна бути лінійною комбінацією інших ознак, інакше вона не несе нової інформації;

нормальний закон розподілу дискримінантних змінних x_{ij} (за ознаками) [16].

У випадку, якщо наведені припущення не задовольняються, то порушується питання доцільності використання дискримінантного аналізу для класифікації нових спостережень [16].

Основні проблеми у використанні ДА [17] наведено на рис. 2.8.

Дискримінантний аналіз може використовуватися і для прогнозування поведінки одиниць статистичної сукупності, що спостерігаються шляхом зіставлення їх з поведінкою аналогічних об'єктів навчальних підмножин [16].

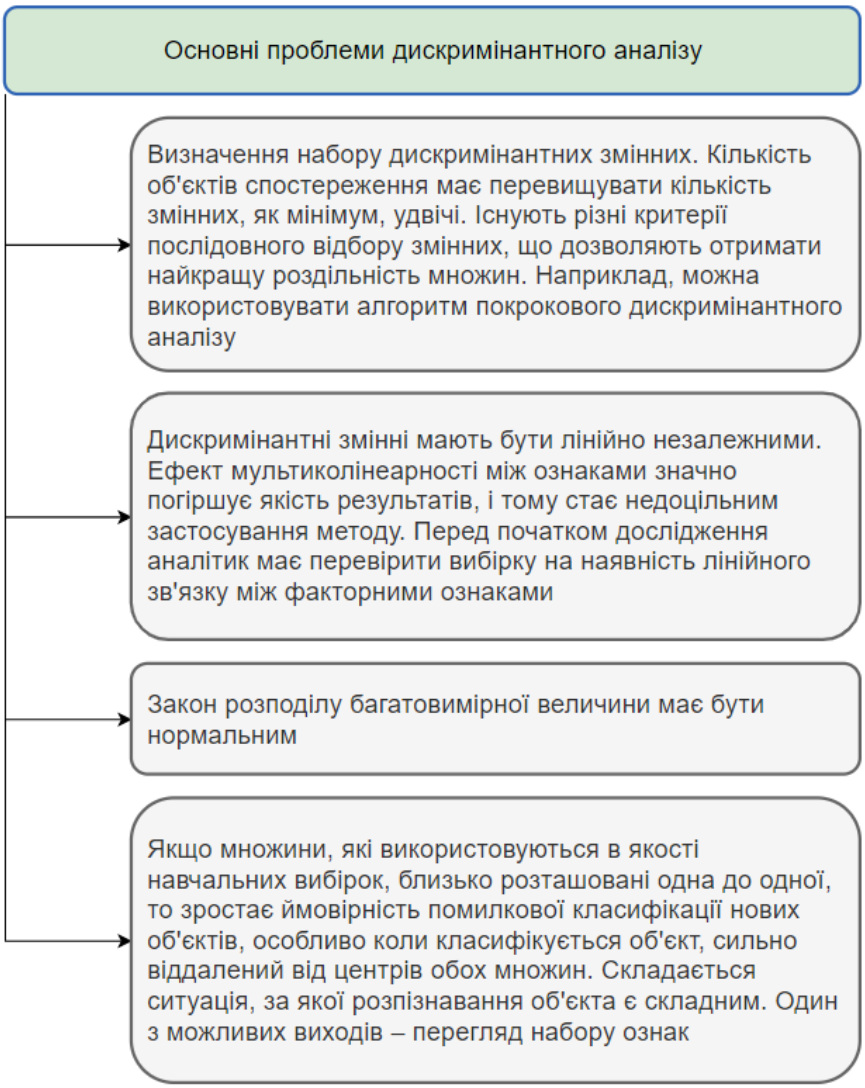

Рис. 2.8. Проблеми використання дискримінантного аналізу

Для практичної реалізації дискримінантного аналізу необхідно мати інформацію про апріорні ймовірності π_j та функції щільності ймовірності $f_j(X)$. Ці значення можуть бути відомими на підставі теоретичних міркувань або результатів попередніх досліджень. Однак, у випадку їх невідомості, їх можна замінити статистичними оцінками, які отримуються на основі доступних навчальних вибірок [18]:

$$\pi_j = \frac{n_j}{n_{sum}},$$

де n_j – обсяг j-ї вибірки;

$n_{sum} = n_1 + n_2 + \cdots + n_k$ – сумарний обсяг навчальних вибірок [18].

Оцінки апріорних ймовірностей, часто позначають як π_j, є важливими параметрами для проведення дискримінантного аналізу. Вони відображають відносну частоту появи кожного класу в навчальній вибірці. У випадку, коли апріорні ймовірності невідомі, їх можна оцінити, наприклад, шляхом підрахунку відносної частоти появи кожного класу в навчальній вибірці [18].

Так само, для проведення дискримінантного аналізу необхідно мати інформацію про функції щільності ймовірності $f_j(X)$, які відображають розподіл ознак у кожному класі. Якщо функції щільності ймовірності невідомі, їх також можна оцінити за допомогою статистичних методів на основі навчальних даних [18].

Отже, для успішної реалізації дискримінантного аналізу важливо мати апріорні ймовірності та функції щільності ймовірності. Якщо ці значення невідомі, їх можна оцінити на підставі статистичних оцінок, отриманих з навчальних вибірок [18].

Щоб оцінити функцій щільності частіше за все використовують 2 підходи (рис. 2.9).

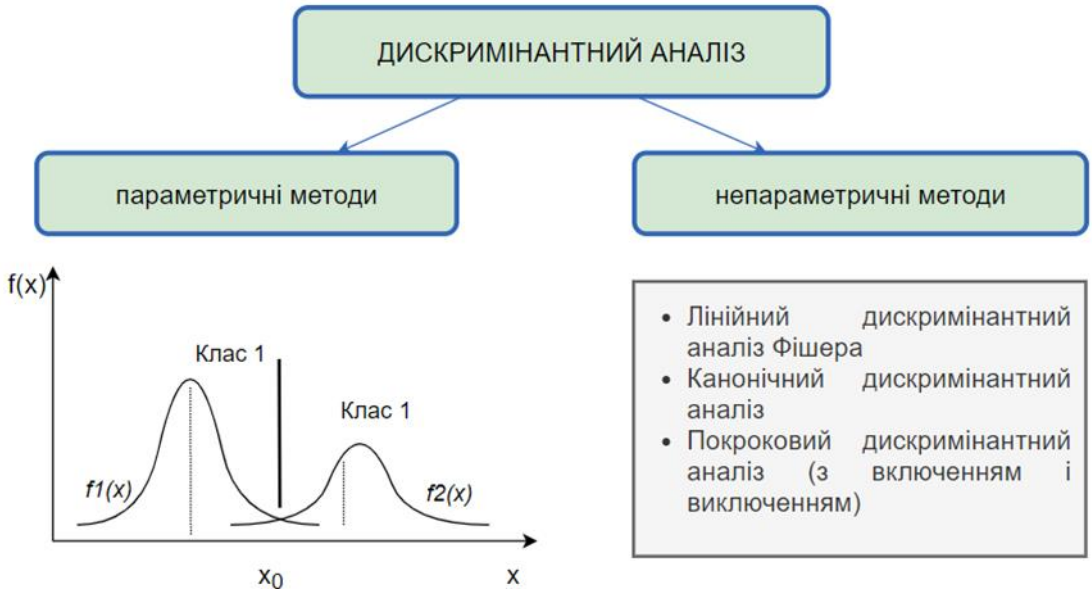

Рис. 2.9. Методи дискримінантного аналізу

У першому підході, який називається параметричним дискримінантним аналізом, припускають, що всі класи мають функції щільності ймовірності, які

належать до однієї параметричної сім'ї $fX(\theta)$ і відрізняються лише значеннями векторного параметра θ. У цьому випадку відповідні значення параметра θ_j оцінюють за спостереженнями, що належать до j-ї вибірки. У другому підході (непараметричний дискримінантний аналіз) загальний вигляд функцій $f_j(X)$ є невідомим. Тому для їх оцінювання необхідно використовувати спеціальні методи, такі як будівництво непараметричних оцінок типу гістограми або ядра [18].

Давайте розглянемо геометричне тлумачення представлення дискримінантних змінних. Проаналізуємо об’єкти, що належать двом різним множинам M_1 і M_2 (рис. 2.10) [18].

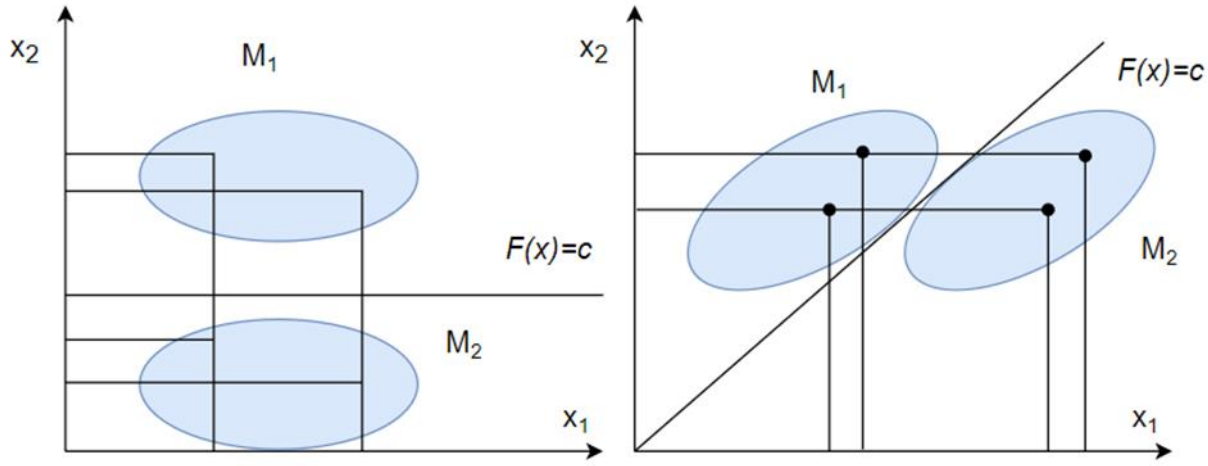

Рис. 2.10. Геометрична інтерпретація дискримінантних функцій і змінних

Як видно з рис.2.10, кожен об’єкт характеризується в даному випадку двома змінними x_1 і x_2. Якщо ми спостерігаємо проекції об'єктів (точок) на кожну вісь, то ці множини перетинаються. Іншими словами, деякі об'єкти, що належать до обох множин, мають подібні характеристики за кожною змінною [18].

Для досягнення оптимального розділення двох заданих множин, необхідно побудувати відповідну лінійну комбінацію змінних x_1 і x_2. У випадку двовимірного простору це завдання полягає в визначенні нової системи координат. Особливістю цієї системи є розташування нових осей L і C таким чином, щоб проекції об'єктів, належних різним множинам, були максимально

розділені на вісі *L*. (рис. 2.11) [18].

Ось *C* перпендикулярна вісі *L* і відділяє дві «множини» точок так, щоб вони опинилися по різним бокам від цієї прямої. Одночасно з цим імовірність помилки класифікації повинна бути мінімальна [18].

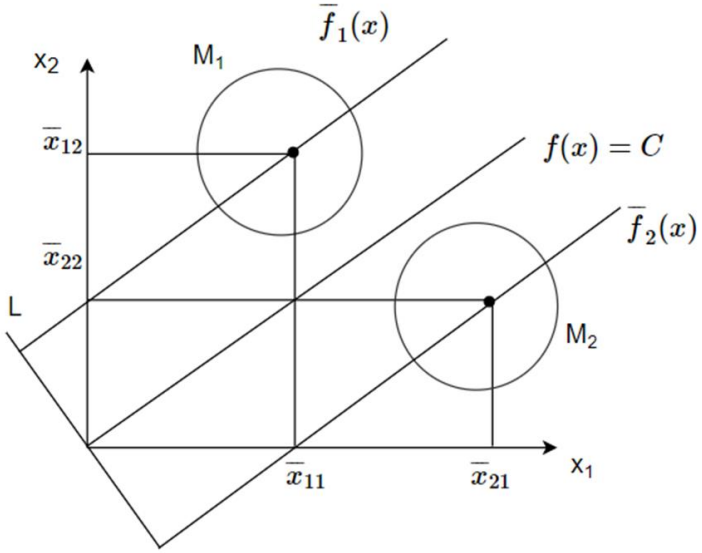

Рис. 2.11. Центри дискримінантних множин і константа дискримінації

Сформульовані умови слід ураховувати під час обчислення коефіцієнтів a_1 і a_2 дискримінатної функції [18]:

$$D = a_1 x_1 + a_2 x_2.$$

Або у загальному випадку:

$$D = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_k x_k.$$

Дискримінантна функція *f(x)* – комбінація показників, коефіцієнти якої підбираються за умови найбільшої різниці функції між відомими класами [18].

Константа дискримінантної функції – межа, що розділяє дві сукупності [18]. Константу обчислюють за формулою:

$$C = \frac{1}{2}(\bar{f_1} + \bar{f_2}).$$

Позначимо через x_{ij} середнє значення j-ї ознаки у об'єктів i-ї множини (класу). Тоді для множини M_1 середнє значення функції $f_1(x)$ буде таким: $\bar{f_1}(x) = a_1(\bar{x}_{11}) + a_2(\bar{x}_{12})$. Для множини M_2 середнє значення функції $f_2(x)$ дорівнюватиме: $\bar{f_1}(x) = a_1(\bar{x}_{21}) + a_2(\bar{x}_{22})$. Геометрична інтерпретація цих функцій включає дві паралельні прямі, які проходять через центри відповідних класів або множин (рис. 2.11) [18].

Коефіцієнти дискримінантної функції аі визначаються так, щоб $f_1(x)$ і $f_2(x)$ найбільше відрізнялися один від одного, тобто щоб для обох множин був максимальний вираз [18]:

$$\overline{f_1(x)} - \overline{f_2(x)} = \sum_{i=1}^{n} a_1 x_{2i} - \sum_{i=1}^{n} a_2 x_{2i}.$$

У цій ситуації поивнна мінімізуватися внутришньогруппова дисперсія – квадрат суми відхилень від середніх значень [18]:

$$\sum_{i=1}^{2} \sum_{t=1}^{n_k} (Y_k - \bar{Y_k}) = A'(X_1'X_1 + X_2'X_2).$$

Разом з тим міжгрупова варіація повинна бути максимальна [18], тобто:

$$(\bar{Y_1} - \bar{Y_2})^2 = A(\overline{X_1} - \overline{X_2})(\overline{X_1} - \overline{X_2})'A.$$

У випадку потреби, можна розбити множину об'єктів на k класів і розрахувати k дискримінантних функцій, щоб відокремити класи один від одного за допомогою індивідуальних роздільних поверхонь (рис. 2.12). Наприклад, якщо ми маємо три сукупності, ми можемо обчислити функцію для дискримінації між сукупністю M_1 та сукупностями M_2 та M_3, які розглядаються

разом, а також іншу функцію для дискримінації між сукупністю M_2 та сукупністю M_3 [18].

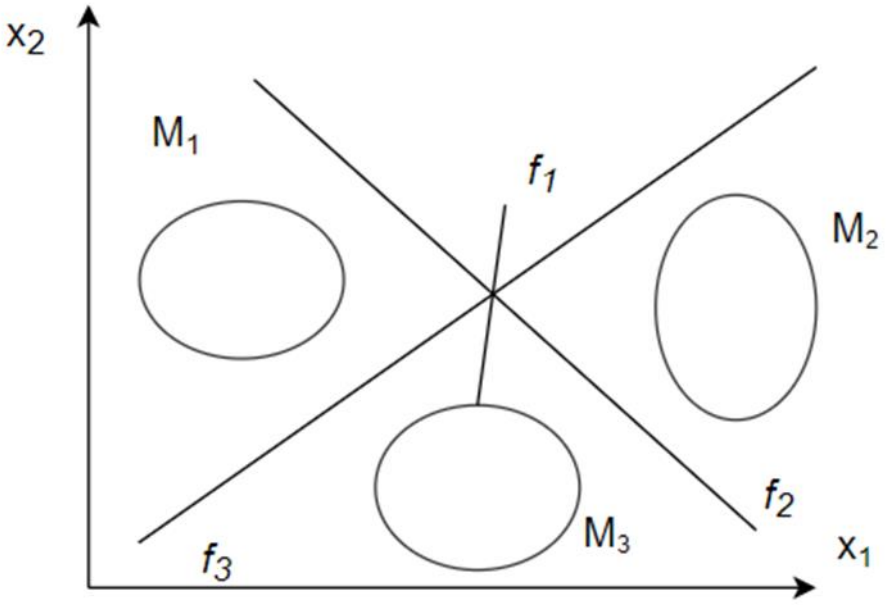

Рис. 2.12. Дискримінантні функції для трьох вибірок

Це підходить для випадків, коли ми бажаємо провести розподіл об'єктів на декілька категорій, і кожна категорія має свої унікальні роздільні поверхні. Цей процес дозволяє використовувати дискримінантні функції для ефективного розподілу об'єктів у відповідні класи [18].

За допомогою дискримінантних функцій ми можемо здійснити відокремлення та класифікацію об'єктів на основі їхніх характеристик та роздільних поверхонь, що дозволяє ефективно проводити аналіз та виявляти взаємні відмінності між різними сукупностями об'єктів [18].

Для оцінювання якості дискримінації між групами можна скористатися критерієм Фішера, відомим як F-критерій. Цей критерій дозволяє визначити оптимальний розподіл груп шляхом максимізації його значення. Застосування критерію Фішера полягає в порівнянні дисперсії міжгрупових відхилень з внутрішньогруповими відхиленнями. Чим більше відхилення між групами (міжгрупова дисперсія) в порівнянні з внутрішньогруповими відхиленнями, тим краще відбувається дискримінація між групами [18].

Отже, шляхом обчислення F-критерію та порівняння його значення з критичним значенням, можна зробити висновок про якість дискримінації та оптимальний розподіл груп. Максимальне значення F-критерію вказує на найкращий розподіл груп та ефективність дискримінації між ними [18]:

$$\frac{A(\overline{X_1} - \overline{X_2})(\overline{X_1} - \overline{X_2})'A}{A'[(n_1 + n_2 - 2)S_*]A} \to max;$$

$$S_* = \frac{1}{n_1 - n_2 - 2}(X_1'X_1 + X_2'X_2).$$

Вектор коефіцієнтів обчислюється таким чином [18]:

$$A' = X_*^{-1}(\overline{X_1} + \overline{X_2}).$$

Розглянемо алгоритм лінійного дискримінантного аналізу Фішера для двох класів за нормального закону розподілу показників [32]. Для успішного використання методу дискримінантного аналізу необхідно враховувати наступні умови:

- Обсяг вибірки повинен бути більшим за кількість змінних. Це гарантує наявність достатньої кількості даних для аналізу та надає статистичну достовірність результатів.
- Кластери, між якими здійснюється дискримінація, повинні підпорядковуватись багатовимірному нормальному розподілу. Це передбачає, що дані в кожному кластері мають близьке до нормального розподілення.
- Класи можуть перетинатися, але їх центри повинні бути достатньо віддаленими один від одного. Це дозволяє ефективно розділяти класи та забезпечує низький рівень перекриття між ними.
- Різниця між коваріаційними матрицями цих кластерів повинна бути статистично незначущою. Це означає, що внутрішні зв'язки в межах кожного

класу подібні та немає суттєвих відмінностей між ними.

- Кількість навчальних вибірок у кластері повинна бути меншою за кількість дискримінантних функцій. Це дозволяє уникнути перенавчання моделі та забезпечити її роботу з ефективним використанням обмежених даних.
- Дотримання цих умов дозволить досягти надійного та точного дискримінантного аналізу для ваших даних [18].

В основу метода лінійного дискримінантного аналізу, запропонованого Р. Фішером у 1936 р., вкладене припущення, що класифікацію можливо виконати за допомоги лінійної комбінації дискримінантних змінних [18].

Нехай є 2 генеральні сукупності X і Y, що мають 3-вимірний закон розподілу з невідомими, але рівними коваріаційними матрицями. З них взяті навчальні вибірки з обсягами n_1 в X і n_2 в Y [18]. Розглянемо алгоритм методу:

1. Задають вхідні матриці X і Y з об'єктами n_1 і n_2 [18], відповідно:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix},$$

$$Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{pmatrix}.$$

2. Записують нове спостереження (z), яке слід класифікувати [18]:

$$Z = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{pmatrix}.$$

3. Обчислюють вектори середніх значень кожної з підмножин [18]:

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} \text{ і } \bar{y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix},$$

$$\bar{x}_j = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{ij}.$$

4. Обчислюють оцінки коваріаційних матриць S_x і S_y [18]:

$$S_x = (S_{ki})_x \text{ і } S_y = (S_{ki})_y.$$

Знаходимо елемент матриці $\overline{S_x}$:

$$S_{ki} = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = \overline{x_j x_k} - \overline{x_j x_k} = 1,2,3,$$

де x_j і x_k – середні значення [18].

5. Визначають незміщену оцінку сумарної коваріаційної матриці [18]:

$$\hat{S} = \frac{1}{n_1+n_2-2}(n_1 S_x + n_2 S_y).$$

6. Визначають матрицю, $\hat{S}^{-1}$, зворотну до матриці $\hat{S}$ [18].

7. Обчислюють вектор оцінок дискримінантної функції [18]:

$$a = \bar{S}^{-1}(\bar{x} - \bar{y}).$$

8. Обчислюють оцінки векторів значень дискриминантної функції для матриць вихідних даних $\widehat{U_x} = Xa$, $\widehat{U_y} = Ya$ [18].

9. Обчислюють середні значення оцінок дискримінантної функції [18]:

$$\overline{\widehat{U_x}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{u_{xi}},\ \overline{\widehat{U_x}} = \frac{1}{n_1}\sum_{i=1}^{n_2}\widehat{u_{yi}}.$$

10. Обчислюють константи [18]:

$$\hat{C} = \frac{1}{2}(\overline{\widehat{u_x}} + \overline{\widehat{u_y}}).$$

11. Записують дискримінантну функцію і її економічну інтерпретацію [18].

12. Для перевірки умови максимально чіткого поділу груп доцільно використовувати критерій «лямбда Уїлкса» [18]:

$$L_W = \frac{1}{\lambda},$$

де λ знаходять відповідно до формули:

$$\lambda = F = \frac{a^T S_{xy} a}{a^T S_* a} \to max.$$

За $L_W \to 1$ групи поділені чітко. За $L_W \to 0$ результати аналізу не можна використовувати [18]. Вплив окремих показників на результати дискримінантного аналізу розраховують так:

$$R_{x_i} = \frac{a_j^*}{\sum_{j=1}^{m} \left|a_j^*\right|},$$

де $\left|a_j^*\right|$ – модуль стандартизованої оцінки показника X_j [18].

13. Обчислюють значення дискримінантної функції для υ-го спостереження, що підлягає дискримінації, вирішивши рівняння [18]:

$$\widehat{U_\upsilon} = z_{\upsilon 1} a_1 + z_{\upsilon 2} a_2 + z_{\upsilon 3} a_3.$$

Правила віднесення об’єктів до класів [18] наведені на рис.2.13:

Рис. 2.13. Правила віднесення об’єктів до класів

Описаний алгоритм можна використовувати у випадку розподілу навчальної вибірки на два класи спостережень [18].

2.3. Когнітивний аналіз та моделювання складних ситуацій

Як відомо, когнітивне моделювання - це метод аналізу, який дозволяє визначити вплив факторів на перетворення об'єкта управління в заданий стан, враховуючи схожість та відмінність у впливі різних факторів на об'єкт управління. [22].

Моделювання стало популярним явищем як у технічних, так і у гуманітарних науках. Основою для розуміння сутності моделювання є поняття "модель" [22].

Модель - це спосіб відображення об'єкта, системи або ідеї у конкретній формі, відмінній від їх реальної форми існування [22].

За когнітивним підходом, процес моделювання можна виразити таким способом (рис. 2.14) [22]:

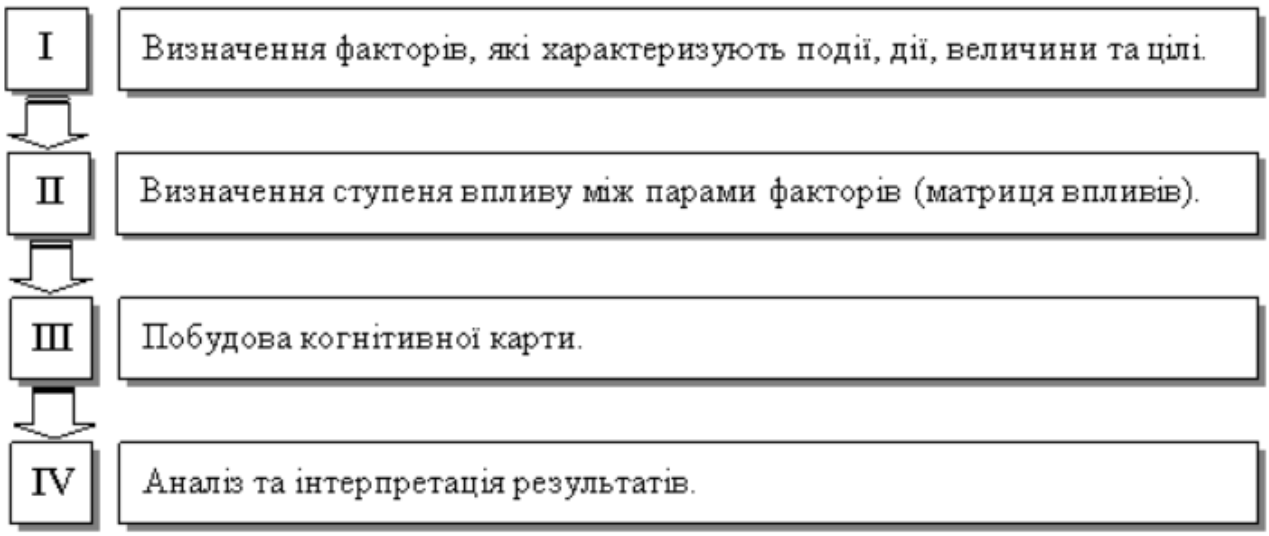

Рис. 2.14. Етапи когнітивного моделювання

Давайте глибше розглянемо деякі з перерахованих етапів, методи виконання конкретних завдань на кожному з цих етапів, а також проблеми, що можуть виникати на різних етапах когнітивного аналізу[22].

Когнітивна карта (від лат. cognitio – знання, пізнання) – образ знайомого просторового оточення [23].

Термін запропонований у 1948 у роботі видатного американського психолога, представника необіхевіоризму Е. Толмена "Когнітивні карти у щурів та людини". Аналізуючи поведінку щурів у лабіринті, Толмен дійшов висновку, що в результаті бігання лабіринтом у щура формується особлива структура, яку можна назвати когнітивною картою навколишнього оточення [23].

Когнітивна карта ситуації є графом з вагами, де [22]:

- Кожна вершина однозначно відповідає базовим факторам ситуації, які описують процеси в цій ситуації.
- Дуги встановлюються безпосередньо через взаємозв'язки між факторами, шляхом аналізу причинно-наслідкових ланцюжків, що описують передачу впливів від одного фактора до іншого [22].

Згідно з прийнятою думкою, фактори, що включені в умову "якщо" у вислові "якщо, ... то", впливають на наслідки "то" цього вислову. Цей вплив може бути позитивним (підсилюючим), негативним (гальмуючим) або залежати від

додаткових умов, які можуть змінювати його знак [22].

У побудові когнітивної моделі виникають дві основні проблеми. По-перше, складності виникають при визначенні факторів (елементів системи) і їх ранжуванні (виділенні основних та вторинних факторів) на етапі побудови орієнтованого графа. По-друге, виникає складність при визначенні ступеня взаємовпливу між факторами (встановлення ваг дуг графа) на етапі побудови функціонального графа [22].

На початку виділяють цільові чинники – ті чинники, зміни яких в потрібну сторону ми хочемо добитися. Їх не повинно бути багато [23]. Далі виділяють важелі дії – ті чинники, які ми можемо в певних межах міняти. Якщо таких немає, то ми лише можемо проаналізувати розвиток ситуації – теж корисне завдання, на зразок прогнозу погоди [23].

Наступними знаходять цикли зворотного зв'язку – тобто замкнуті шляхи на графі. Ці цикли можуть бути такими, що підсилюють відхилення і що стабілізують – щоб це дізнатися, треба перемножити всі знаки ребер шляху. Якщо вийшов «+», то цикл є підсилюючим, а якщо «–», то що стабілізуючим [23].

Аналізують зв'язки (не тільки прямі) важелів дії і цільових чинників – як взагалі ми можемо управляти ситуацією [23].

Когнітивна карта обмежується лише відображенням наявності впливів між чинниками, не уточнюючи деталей цих впливів, динаміки зміни впливів або тимчасових змін чинників. Для врахування всіх цих аспектів потрібно перейти до когнітивної моделі, яка наступним рівнем структуризує інформацію, що міститься в когнітивній карті. У когнітивній моделі кожен зв'язок між чинниками розкривається до відповідного рівняння, яке включає як кількісні, так і якісні змінні. Формально, когнітивна модель може бути графом, де кожна дуга відображає функціональну залежність між чинниками, створюючи функціональний граф моделі ситуації [24].

Розглянемо можливі математичні інтерпретації когнітивних карт:

1. М'які математичні моделі.

Основна особливість м'яких математичних моделей полягає в тому, що

вони допускають введення нечітких, або нестрогих, критеріїв і правил, що відображають невизначеність, неоднозначність або нечіткість інформації, яка часто присутня в людському мисленні. Це дозволяє враховувати невизначеність в процесі моделювання і отримувати більш реалістичні результати. [23].

Перевагами цих методів є: врахування нечіткості – м'які математичні моделі дозволяють враховувати нечіткість та неоднозначність даних і правил, що дозволяє більш точно відтворити складність людського мислення і прийняття рішень; гнучкість – вони можуть бути адаптовані до зміни умов і нової інформації, що робить їх більш гнучкими в порівнянні зі стандартними математичними моделями; універсальність – м'які математичні моделі можуть бути застосовані в різних галузях, включаючи науку, технології, управління, економіку і соціальні науки. [23].

Істотним мінусом є те, що математика працює все-таки з достатньо простими моделями. Якщо система достатньо складна, то описати всі її можливі рішення математика безсила, можливо тільки чисельне моделювання [23].

2. Модель підсумовування впливу факторів.

Зазвичай взаємодія факторів не має конкретного реального механізму і описується словесно, нечітко. Часто взаємодію факторів описують експерти таким чином: "При значному зростанні фактора А, фактор Б незначно зменшується". Оскільки немає конкретних одиниць вимірювання, ми можемо сформулювати закон виду: "Якщо значення фактора k зростає на X_k відсотків, то значення фактора m зменшується на X_m відсотків". Це можна виразити у формулі. [22]:

$$X_m(t+1) = W_{m,k} \cdot X_k(t),$$

де $X_m(t+1)$ – значення фактору у наступний момент часу;

$W_{m,k}$ – коефіцієнт зміни фактора;

$X_k(t)$ – зростання фактору [22].

Усі взаємодії між факторами моделі визначаються за допомогою матриці суміжності (також відомої як матриця впливів) вершин орієнтованого графа,

позначеної як - $W = (W_{m,k})$. Якщо на когнітивній карті відсутнє ребро, що з'єднує вершину k з вершиною m, тоді $W_{m,k} = 0$. Фактично, кожному ребру графа, крім знаку, присвоюється його вага [22].

Зазвичай вимагають, щоб значення $W_{m,k}$ знаходилося в діапазоні від -1 до 1. Це відповідає тому, що розглядається система з інерцією, де зміна одного фактора має незначний вплив на зміни інших факторів. [22].

Для подальшого аналізу необхідно розглянути модель колективного впливу декількох зв'язків на фактор. Якщо до однієї вершини входить кілька стрілок, необхідно вивчити, як взаємодіють зміни, що відбуваються по кожній стрілці. Усі взаємодії між змінами факторів в момент часу $t + 1$ визначаються матрицею суміжності W орієнтованого графа та вектором змін факторів в момент часу t. [22]:

$$X(t+1) = F(X(t), W),$$

де $X(t+1)$ – значення фактору в наступний період часу;

$F(X(t), W)$ – функція впливу матриці суміжності [22].

Часто використовується більш проста інтерпретація, яка базується на виконанні операції сумування [22]:

$$X(t+1) = W \cdot X(t),$$

де $X(t+1)$ – значення фактору у наступний момент часу;

W – матриця суміжності;

$X(t)$ – значення фактору у момент часу t [22].

Вигляд $X(t+1) = F(X(t), W)$ через операцію суми [22]:

$$X_m(t+1) = \sum_k W_{m,k} \cdot X_k(t),$$

де $X_m(t+1)$ – значення фактору у наступний момент часу;

$W_{m,k}$ – коефіцієнт зміни фактора;

$X_k(t)$ – зростання значення фактору [22].

3. Модель нелінійної взаємодії факторів.

Модель враховує вплив усіх діючих факторів, але при цьому керується найсильнішим з них. Цей принцип можна інтерпретувати ще одним способом - коли експерт оцінює силу впливу однієї причини на наслідок, припускаючи, що інші фактори не мають впливу. У реальності це не завжди вірно. Оцінка сили впливу враховує певний сумарний результат всіх причин при умові, що інші причини є незначними. Тоді $X_m(t+1) = \sum_k W_{m,k} \cdot X_k(t)$ приймає вигляд [22]:

$$X_m(t+1) = W_{m,N} \cdot X_N(t),$$

де N – таке k, при якому досягається $max_k\left(\left|W_{m,k} \cdot X_k(t)\right|\right)$,

$X_m(t+1)$ – значення фактору у наступний момент часу;

$W_{m,k}$ – коефіцієнт зміни фактора;

$X_k(t)$ – зростання значення фактору [22].

4. Нечітка модель взаємодії факторів.

Часто врахування взаємодії між факторами засноване на нечітких, приблизних міркуваннях. У таких випадках висновки також є приблизними. Для оцінки достовірності цих висновків використовуються механізми нечіткої логіки. Науковці пропонують механізм оцінки достовірності отриманих висновків, що базується на побудованій моделі. Для цього обчислюється значення консонансу за певними формулами [22]:

$$C_m(t) = \frac{|z_m^+(t) + z_m^-(t)|}{|z_m^+(t)| + |z_m^-(t)|},$$

$$z_m^+(t) = max_k\left(W_{m,k} \cdot X_k(t)\right),$$

$$z_m^-(t) = max_k\left(-W_{m,k} \cdot X_k(t)\right),$$

де $z_m^-(t)$, $z_m^+(t)$ – дія на фактор;

$W_{m,k}$ – коефіцієнт зміни фактора;

$X_k(t)$ – зростання фактору [22].

Значення консонансу відображає рівень впевненості в отриманому висновку і його відповідність очікуваній інформації. Чим вище значення консонансу, тим більш точним вважається висновок. Максимальний рівень впевненості (рівний 1) досягається, коли відсутні фактори, що впливають у протилежних напрямках. Мінімальний рівень (рівний 0) спостерігається, коли існують фактори, що мають приблизно однаково сильний, але протилежний вплив [22].

Варто відзначити, що запропонований метод оцінки консонансу може бути застосований і для ситуацій лінійного характеру $X_m(t+1) = \sum_k W_{m,k} \cdot X_k(t)$. Тоді для обчислення значення консонансу слід використовувати наступну формулу [22]:

$$C_m(t) = \frac{\left|\sum_k W_{m,k} \cdot X_k(t)\right|}{\sum_k \left|W_{m,k} \cdot X_k(t)\right|},$$

де $W_{m,k}$ – коефіцієнт зміни фактора;

$X_k(t)$ – зростання фактору [22].

Інтервали значень консонансу можуть бути лінгвістично інтерпретовані у вигляді таких термінів, як "неможливо", "можливо", "достовірно" і тому подібне [22].

Узагальнюючи, когнітивне моделювання [33] є важливим інструментом у вивченні когнітивних процесів, а також має широкий спектр застосувань у наукових, технологічних та практичних галузях. Воно допомагає нам краще зрозуміти людську природу та розробляти нові інноваційні рішення, що базуються на наших пізнаннях про розумову діяльність.

2.4. Засоби реалізації методів кластерного та дискримінантного аналізів, а також побудови когнітивної моделі

Як відомо, найпопулярнішими засобами реалізації методів кластерного і дискримінантного аналізів є їх реалізація у Microsoft Excel, за допомоги мови програмування R [34] та у середовищі STATISTICA.

У власній роботі я вирішив виконати реалізацію кластерного аналізу у середовищі RStudio, а реалізацію дискримінантного аналізу – за допомоги ППП STATISTICA.

STATISTICA – це система для статистичного аналізу даних, що включає широкий набір аналітичних процедур і методів: більше 100 різних типів графіків, описові та внутрішньогрупові статистики, розвідувальний аналіз даних, кореляції, швидкі основні статистики та блокові статистики, інтерактивний імовірнісний калькулятор, Т-критерії (і інші критерії групових відмінностей), таблиці частот, сполученості, прапорів і заголовків, аналіз багатовимірних відгуків, множинна регресія, непараметричні статистики, загальна модель дисперсійного та коваріаційного аналізу, підганяння розподілів, видобуток даних, нейронні мережі та багато іншого [25].

Продукти серії STATISTICA засновані на найсучасніших технологіях, повністю відповідають останнім досягненням в галузі IT, дозволяють вирішувати будь-які завдання в галузі аналізу та обробки даних, ідеально підходять для вирішення практичних завдань у маркетингу, фінансах, страхуванні, економіці, бізнесі, промисловості, медицині і т.д [25].

Система STATISTICA надає користувачам усіх областей унікальні можливості щодо аналізу даних. Система розрахована не так на математиків, але в фахівців у прикладних областях [25].

R - це мова програмування та середовище розробки, спеціально призначені для статистичного аналізу, візуалізації даних та машинного навчання. Вона є однією з найпопулярніших мов у наукових та статистичних галузях і використовується для обробки та аналізу даних, створення статистичних

моделей, виконання експериментів, візуалізації результатів та багато іншого. [26]. Тож, чому саме R?

У R є багато переваг, які дозволяють використовувати саме цю програму [26]:

1. Безкоштовність та відкритий вихідний код: R є вільним програмним забезпеченням, що означає, що ви можете використовувати, модифікувати та поширювати його безкоштовно. Відкритий вихідний код сприяє активному співробітництву спільноти та розвитку багатофункціональних пакетів.

2. Потужність та гнучкість: R надає широкий спектр функцій та пакетів для статистичного аналізу, обробки даних, машинного навчання та візуалізації. Вона має багатий набір статистичних методів і алгоритмів, які дозволяють вам виконувати складні аналізи та моделювання.

3. Велика спільнота користувачів та підтримка: R має активну та розширену спільноту користувачів, яка надає допомогу, навчальні матеріали, пакети та розв'язання проблем. Це означає, що ви можете легко знайти підтримку та ресурси для вирішення своїх завдань.

4. Візуалізація даних: R має потужні засоби для графічної візуалізації даних. Ви можете побудувати різноманітні графіки, діаграми та візуалізації для вивчення та представлення ваших даних.

5. Інтеграція та робота з іншими мовами програмування: R може взаємодіяти з іншими мовами програмування, такими як Python, Java, C ++ та іншими. Це дозволяє використовувати R для аналізу даних та статистики разом з іншими інструментами та бібліотеками [26].

На сьогоднішній день R є неперевершеним лідером серед вільно поширюваних систем статистичного аналізу. Це підтверджується такими фактами, як перемога системи R в 2010 році в престижному конкурсі Bossie Awards у кількох категоріях. Вона широко використовується провідними університетами світу, аналітиками великих компаній та наукових центрів для науково-технічних обчислень та реалізації великих інформаційних проектів. Завдяки широкому застосуванню пакетів R для статистичного аналізу і великій

підтримці наукового співтовариства, скрипти R поступово стають визнаним "стандартом" як для наукових публікацій, так і для неформального обміну дослідниками по всьому світу [26].

Для створення когнітивної моделі я використовував середовище Microsoft Excel [35]. Крім цього, наявні й інші діалогові комплекси, які призначені для підтримки процесу прийняття рішень та реалізації когнітивного моделювання, такі як «КОМПАС», «КАНВА», «СИТУАЦІЯ-2» [22].

«КОМПАС» є інструментом, що допомагає в процесі прийняття рішень та здійсненні когнітивного моделювання для широкого спектру ситуацій, що виникають у галузях економіки, політики, соціології, менеджменту та державного управління. Ця програма дозволяє обробляти та систематизувати експертні міркування щодо проблемних ситуацій [22].

«КАНВА» використовується для аналізу та моделювання складних економічних, соціальних та політичних ситуацій. Цей інструмент призначений для розробки стратегій управління та механізмів їх реалізації, а також для створення програмних документів стратегічного розвитку країни, регіону, підприємства, фірми тощо. Крім того, «Канва» використовується як інструмент для безперервного моніторингу стану ситуації, пошуку та перевірки гіпотез щодо механізмів розвитку та управління ситуацією [22].

Для підтримки технології когнітивного моделювання існує діалоговий комплекс під назвою «Ситуация-2». Цей комплекс дозволяє швидко, комплексно і системно охарактеризувати і обґрунтувати сформовану ситуацію навіть в складних і невизначених умовах. Завдяки його функціональності, на якісному рівні можуть бути запропоновані шляхи вирішення проблем, які виникли [22].

Варто відзначити, що ці програмні продукти переважно доступні у безкоштовних демо-версіях, які мають певні функціональні обмеження. Крім того, значним недоліком є складність інтерфейсу деяких програм, що обмежує коло потенційних користувачів [22].

У цей же час ми можемо виокремити кілька переваг для побудови когнітивних моделей саме в Microsoft Excel [36]:

1. Простота використання: Excel є дуже поширеним і зрозумілим інструментом для багатьох користувачів. Він має інтуїтивний інтерфейс, що дозволяє легко створювати та редагувати дані та формули.

2. Таблична організація даних: Excel базується на табличній структурі, що дозволяє зручно організовувати дані та виконувати розрахунки. Це корисно для моделювання когнітивних процесів, оскільки можна представити різні змінні, параметри та залежності у вигляді таблиць.

3. Функціональні можливості: Excel має широкий спектр вбудованих функцій і формул, що дозволяють виконувати обчислення, статистичний аналіз та інші операції. Це дозволяє нам виконувати складні розрахунки та аналізувати дані, що використовуються у когнітивних моделях.

4. Гнучкість: Excel надає можливість швидко змінювати дані та формули, що дозволяє проводити експерименти та тестувати різні гіпотези. Ми можемо вносити зміни в модель, оновлювати дані та отримувати миттєві результати.

5. Візуалізація даних: Excel має різні інструменти для візуалізації даних, такі як діаграми, гістограми та графіки. Це дозволяє зробити когнітивну модель більш доступною та зрозумілою, відображаючи результати та тренди у зрозумілій формі.

Враховуючи ці переваги, Microsoft Excel може бути зручним інструментом для швидкого прототипування, тестування та аналізу когнітивних моделей.

РОЗДІЛ 3. РОЗРОБКА І РЕАЛІЗАЦІЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ РЕЗУЛЬТАТУ ФУТБОЛЬНОГО МАТЧУ ТА АНАЛІЗУ ЯКОСТІ ГРИ КОМАНД

3.1. Розподіл вхідних даних за допомогою методів кластерного аналізу у середовищі RStudio

Для реалізації моделі потрібно спочатку завантажити пакети даних із бібліотеки R. Оскільки ми рахуємо кластерний аналіз, тому нам необхідно завантажити низку пакетів даних через install.packages. А саме, нам потрібно заздалегідь використати пакети psych, factoextra, cluster, NbClust, clValid, ggplot2, MASS, gplots [37].

Наведемо короткий опис обраних бібліотек для аналізу та візуалізації даних:

1. Psych [38]: Бібліотека psych в R надає функції для психометричного аналізу та вимірювання. Вона містить інструменти для визначення надійності, валідності, факторного аналізу, аналізу кластерів та інших психологічних досліджень.

2. Factoextra [39]: Бібліотека factoextra дозволяє виконувати розширений аналіз факторів (principal component analysis) та візуалізувати його результати. Вона надає функції для відображення факторних навантажень, графіків внеску, кругових діаграм факторів та багато іншого.

3. Cluster [40]: Бібліотека cluster містить алгоритми кластеризації для групування схожих об'єктів у кластери. Вона надає методи, такі як k-means, hierarchical clustering, fuzzy clustering та інші. Ця бібліотека допомагає візуалізувати результати кластеризації та оцінити якість кластеризації.

4. NbClust [41]: Бібліотека NbClust надає набір індексів та методів для визначення оптимальної кількості кластерів у даних. Вона використовує різні критерії, такі як коефіцієнт силуету, індекс Хопкінса та багато інших, щоб допомогти визначити найкращу кількість кластерів у кластеризаційних аналізах.

5. clValid [42]: Бібліотека clValid надає інструменти для валідації кластерних аналізів. Вона містить функції для обчислення різних індексів валідності кластерів, таких як індекс Dunn, індекс Silhouette, індекс Ratkowsky-Lance та багато інших. Ці індекси надають числову оцінку якості кластерів, враховуючи внутрішню та зовнішню подібність кластерів.

6. ggplot2 [43]: Бібліотека ggplot2 є потужним інструментом для візуалізації даних. Вона базується на концепції "граматики графіків" і дозволяє створювати високоякісні графіки з великою гнучкістю та контролем. ggplot2 пропонує широкий спектр графічних типів, можливостей налаштування вигляду графіків та можливостей фасетування даних.

7. MASS [44]: Бібліотека MASS (Modern Applied Statistics with S) надає широкий спектр функцій для статистичного аналізу даних. Вона містить методи для лінійної регресії, логістичної регресії, дисперсійного аналізу, кластерного аналізу, довірчих інтервалів та багато інших. MASS також має кілька наборів даних для демонстрації методів та алгоритмів.

8. Gplots [45]: Бібліотека gplots надає різноманітні функції для візуалізації даних та графіків. Вона включає інструменти для створення графіків розсіювання, гістограм, діаграм боксів, тривимірних графіків, класифікаційних матриць та інших типів візуалізацій.

Отже, встановимо та запустимо ці бібліотеки, прописавши відповідний код (рис. 3.1):

```
Diplom.R*
Source on Save                                                    Run
 1  install.packages("psych")
 2  install.packages("factoextra")
 3  install.packages("cluster")
 4  install.packages("NbClust")
 5  install.packages("clValid")
 6  install.packages("ggplot2")
 7  install.packages("MASS")
 8  install.packages("gplots")
 9  #Завантаження пакетів для реалізації кластерного аналізу роботи програми
10  library(psych)
11  library(factoextra)
12  library(cluster)
13  library(NbClust)
14  library(clValid)
15  library(ggplot2)
16  library(MASS)
17  library(gplots)
```

Рис. 3.1. Завантаження та запуск потрібних бібліотек

Наступним кроком є завантаження зібраних статистичних даних усіх команд (окрім двох, між якими будемо пізніше будувати прогноз) за «Чемпіонатом Англійської Премьєр Ліги сезону 2022-2023» [46].

Зробимо це за допомоги Import Dataset → From Excel (рис. 3.2):

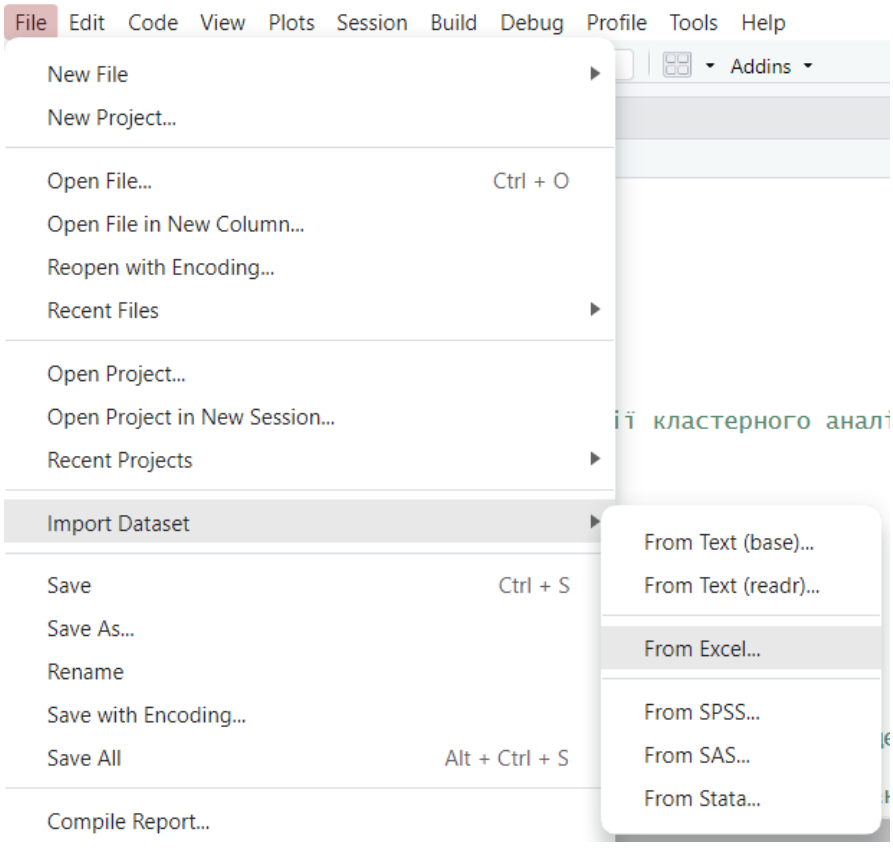

Рис. 3.2. Спосіб завантаження даних з Excel в RStudio

Завантаживши, можемо тепер перевірити наші дані в середовищі (рис. 3.3) за допомогою функції View(football), де football – назва підвантаженого файла.

Diplom.R* × | football ×

Filter

	Назва команди	Відсоток доступних ключових гравців	Кількість перемог	Кількість нічиїх	Відсоток програних матчів	Середня кількість забитих голів	Середня кількість пропущених голів
1	Манчестер Сіті	0.9600	28	4	0.1111111	2.5555556	0.8611111
2	Арсенал	0.9200	25	6	0.1388889	2.3055556	1.1666667
3	Ньюкасл Юнайтед	0.9500	19	12	0.1388889	2.0277778	0.8611111
4	Ліверпуль	0.9150	19	8	0.2500000	1.9444444	1.1666667
5	Брайтон	0.9200	17	7	0.3333333	1.8333333	1.2500000
6	Тоттенгем	0.8960	17	6	0.3611111	1.8055556	1.6388889
7	Астон Вілла	0.8560	17	6	0.3611111	1.3333333	1.2222222
8	Брентфорд	0.9150	16	9	0.3055556	1.5000000	1.2500000
9	Фулгем	0.8560	15	6	0.4166667	1.4444444	1.3611111
10	Челсі	0.8546	11	10	0.4166667	1.0000000	1.1388889
11	Кристал Пелес	0.9150	11	10	0.4166667	1.0277778	1.2777778
12	Борнмут	0.8120	11	6	0.5277778	1.0277778	1.9166667
13	Вест Гем	0.8500	10	7	0.5277778	1.0555556	1.4444444
14	Ноттінгем Форест	0.8365	8	10	0.5000000	1.0000000	1.8611111
15	Евертон	0.8465	7	11	0.5000000	0.8888889	1.5555556
16	Лідс	0.8460	7	10	0.5277778	1.2777778	1.9722222
17	Лестер	0.8790	8	6	0.6111111	1.3611111	1.8611111
18	Саутгемптон	0.8933	6	6	0.6666667	0.8611111	1.8333333

Рис. 3.3. Вхідні дані

Далі буде логічно зробити наш датасет у форматі dataframe: football <- as.data.frame(football). Це робими для того, щоб не вийшло так, що наші дані умовно «розсипались» і давали нам неправдиві оцінки. Саме тому візьмемо наш вхідний датасет у форматі dataframe [27].

Наступною нашою дією є стандартизація даних. Оскільки ми використовуємо дані, які мають різну структуру в своєму складі, то нам потрібно провести стандартизацію, що дозволить привести наші початкові дані до однієї системи координат, тобто, до однієї метрики: football_scale <- scale(football[-1], center=T, scale=T). Результат можемо побачити на рис. 3.4:

Diplom.R × football_scale × football ×

Filter

	Відсоток доступних ключових гравців	Кількість перемог	Кількість нічиїх	Відсоток програних матчів	Середня кількість забитих голів	Середня кількість пропущених голів
1	1.8053240	2.2201338	-1.66887609	-1.7574942	2.14767594	-1.59097575
2	0.8489313	1.7443908	-0.78535345	-1.5855654	1.65833206	-0.72792589
3	1.5662259	0.7929049	1.86521445	-1.5855654	1.11461663	-1.59097575
4	0.7293823	0.7929049	0.09816918	-0.8978503	0.95150200	-0.72792589
5	0.8489313	0.4757430	-0.34359214	-0.3820640	0.73401583	-0.49254866
6	0.2750957	0.4757430	-0.78535345	-0.2101352	0.67964429	0.60587844
7	-0.6812970	0.4757430	-0.78535345	-0.2101352	-0.24467194	-0.57100773
8	0.7293823	0.3171620	0.53993050	-0.5539927	0.08155731	-0.49254866
9	-0.6812970	0.1585810	-0.78535345	0.1337224	-0.02718577	-0.17871234
10	-0.7147707	-0.4757430	0.98169182	0.1337224	-0.89713046	-0.80638497
11	0.7293823	-0.4757430	0.98169182	0.1337224	-0.84275891	-0.41408958
12	-1.7333289	-0.4757430	-0.78535345	0.8214375	-0.84275891	1.39046921
13	-0.8247559	-0.6343239	-0.34359214	0.8214375	-0.78838737	0.05666489
14	-1.1475384	-0.9514859	0.98169182	0.6495087	-0.89713046	1.23355106
15	-0.9084402	-1.1100669	1.42345313	0.6495087	-1.11461663	0.37050120
16	-0.9203951	-1.1100669	0.98169182	0.8214375	-0.35341503	1.54738737
17	-0.1313712	-0.9514859	-0.78535345	1.3372238	-0.19030040	1.23355106
18	0.2105392	-1.2686479	-0.78535345	1.6810814	-1.16898817	1.15509198

Рис. 3.4. Стандартизовані дані

Маючи ці результати, переходимо до наступного етапу побудови кластерного аналізу, а саме, побудови дендограми. Для цього скористаємося методом Уорда, для того, щоб показати найбільш правильну дендограму нашого дослідження [27]:

football_res <- hclust(dist(football_scale, method = "euclidean"), method = "ward.D2") #запускаємо алгоритм Уорда

graphs_vision <- cutree(football_res, k = 3) # отримали 3 візуальні кластери

plot(football_res, cex = 0.8) # побудова графіку

rect.hclust(football_res, k = 3, border = 2:4) # візуально підтверджується розбиття на 2 кластери, але так як нам треба розподілити команди на 3 групи, то робимо 3 кластери і додаємо рамки

Так, як для нашої моделі потрібно розбити дані на 3 умовних кластери

(перший – команди переможці, другий – нічия, третій – поразка), то значення k буде дорівнювати 3. Результат побудови дендограми можемо побачити на рис.3.5:

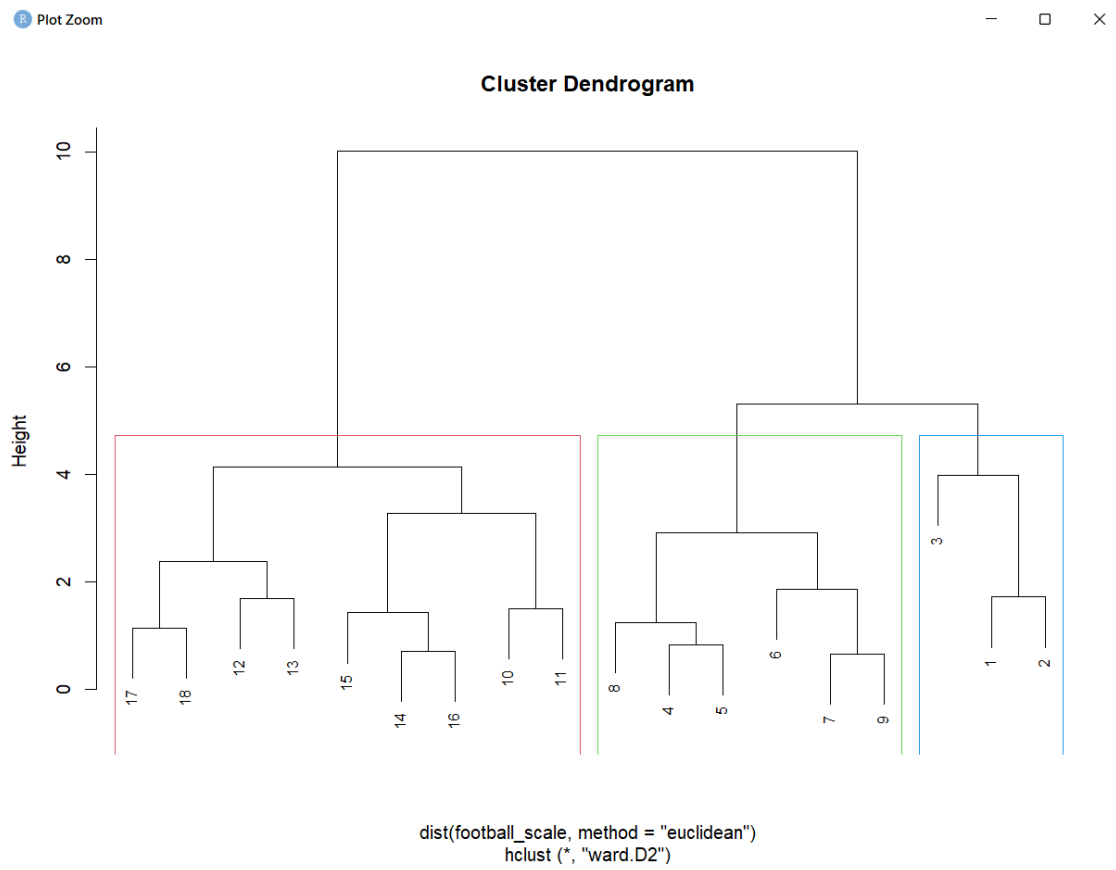

Рис. 3.5. Дендограма дослідження

Далі подивимось, які варіанти кількості кластерів нам можуть запропонувати алгоритми перевірки оптимальної кількості кластерів. В рамках R, ми можемо перевіряти оптимальну кількість кластерів за допомогою трьох вбудованих методів, таких як:

1. Автоматична перевірка методом "ліктя" – fviz_nbclust(football_scale, kmeans, method = "wss") (рис. 3.6):

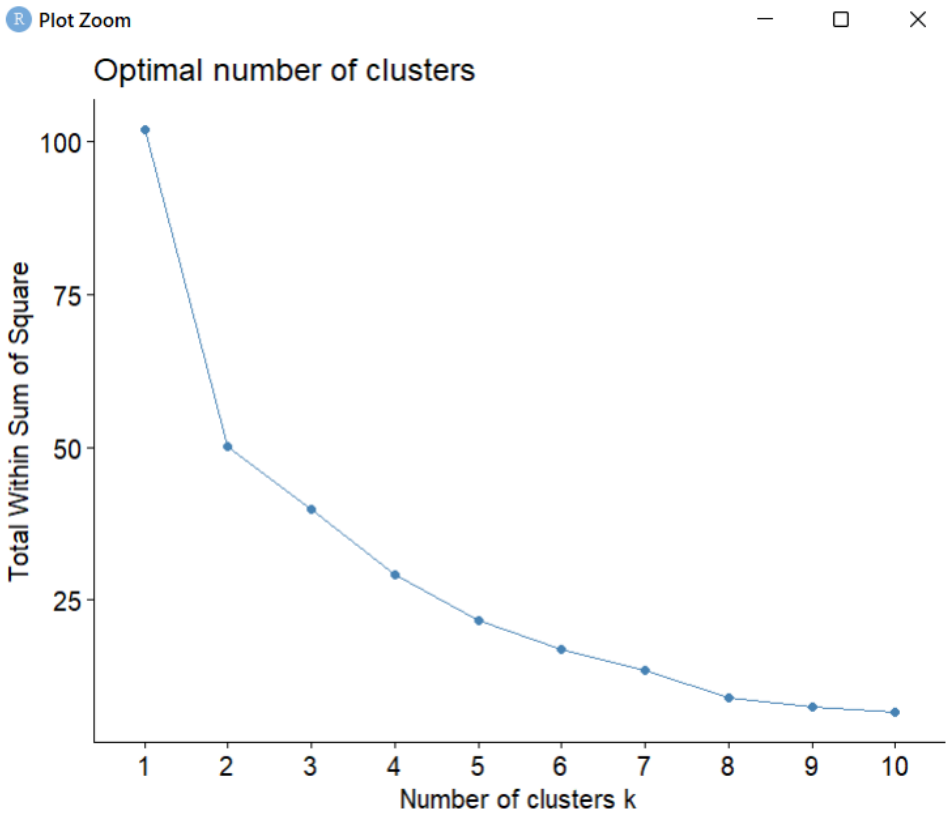

Рис. 3.6. Перевірка оптимальної кількості кластерів методом «ліктя»

Цей метод показує фактично графік кам'янистого осипу, де ми маємо з'ясувати, де у нас відбувається максимально пологий спуск. І як ми можемо побачити, в 2 факторі уже спуск є більш пологий, після 3 спуск стає ще більш пологий. В принципі, за результатами можемо сказати про оптимальність кластерів у кількості двох або трьох.

2. Автоматична перевірка силуетним методом – fviz_nbclust(football_scale, kmeans, method = "silhouette") (рис. 3.7):

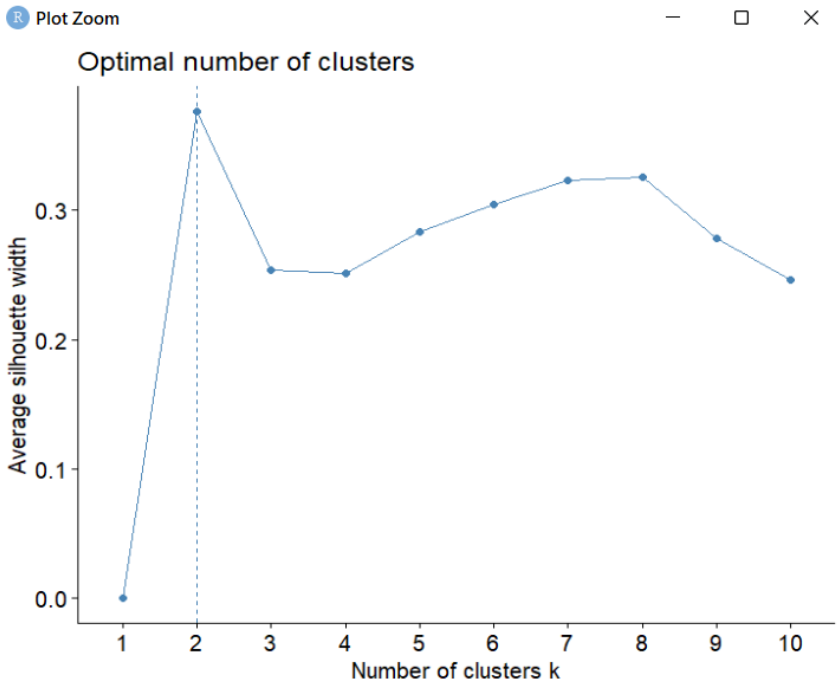

Рис. 3.7. Перевірка оптимальної кількості кластерів силуетним методом

Силуетний метод показує нам, що варто користуватися нам розподілом на 2 кластери.

3. Автоматична перевірка методом gap-статистики – fviz_nbclust(football_scale, kmeans, method = "gap_stat") (рис. 3.8):

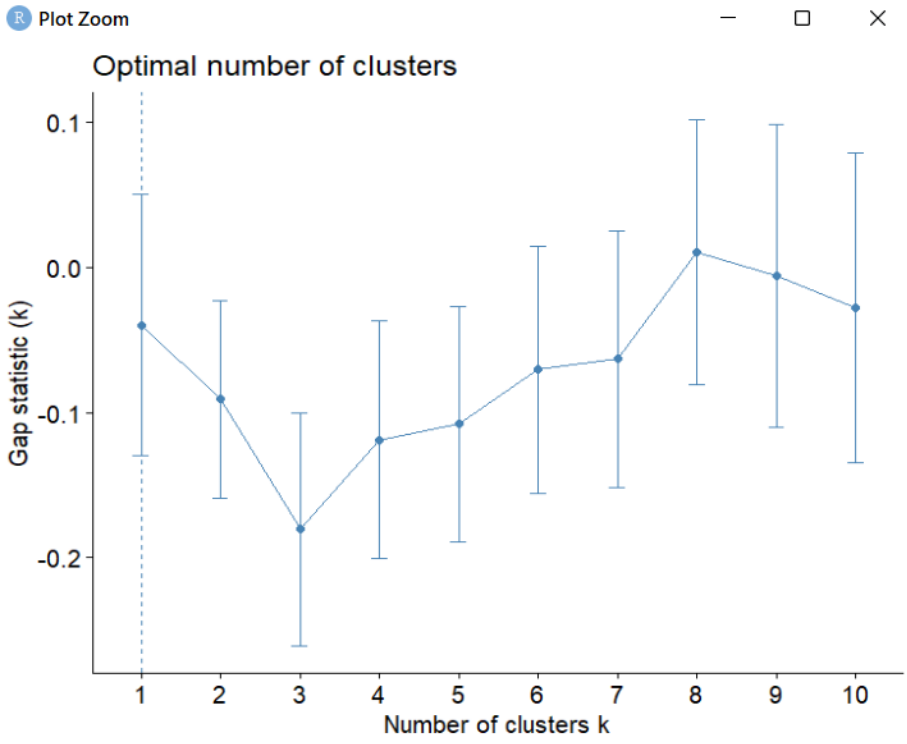

Рис. 3.8. Перевірка оптимальної кількості кластерів методом gap-статистики

Метод gap-статистики рекомендує використовувати нам взагалі лише 1 кластер для побудови моделі, що є нелогічним.

Отже, проаналізувавши усі методи і звевши їх до нашої умови, я обрав розподіл саме на 3 кластери, що буде корисним і правильним рішенням для подольших побудов моделей.

Тепер, знаючи кількість кластерів, яка нам необхідна, переходимо до реалізації метода к-середніх:

```
>football_kmeans <- kmeans(football_scale, 3, nstart = 18)
>fviz_cluster(football_kmeans, data = football_scale, frame.type = "convex")+theme_minimal()
```

Запустивши першу строку коду у нас відбувається розрахунок кластерного аналізу і за допомоги другої строки ми можемо візуально побачити, як графічно можна представити ці величини (рис. 3.9):

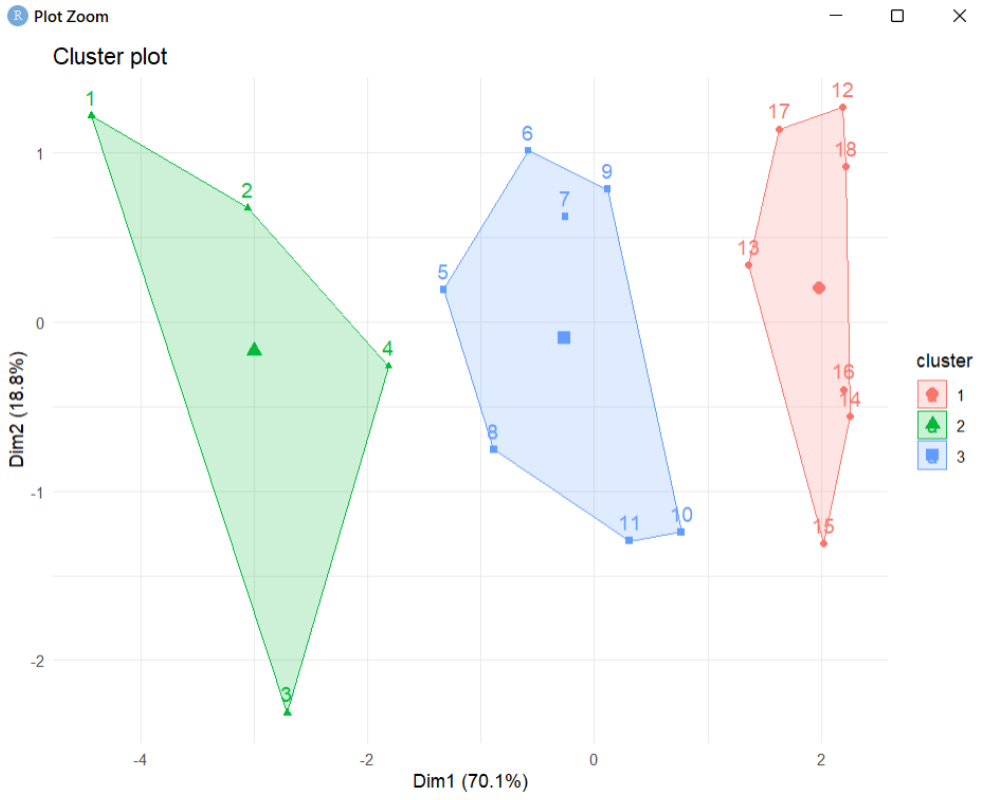

Рис. 3.9. Графічне представлення утворених кластерів

Отже, ми отримали графік трьох кластерів: зелений – команди переможці,

синій – команди, шанс яких зіграти внічию найвищий, та червоний – команди, які мають високий шанс на поразку.

На наступному етапі у нас є змога подивитися поелементне входження наших команд до кластерного аналізу. Для цього застосуємо такі функції:

> View(football[football_kmeans$cluster==1,])

> View(football[football_kmeans$cluster==2,])

> View(football[football_kmeans$cluster==3,])

У результаті виконання першої строки ми отримаємо ті команди, які увійшли до групи команд «лузерів» (рис. 3.10):

Diplom.R | football[football_kmeans$cluster == 1,] | football_scale | football

Filter

	Назва команди	Відсоток доступних ключових гравців	Кількість перемог	Кількість нічиїх	Відсоток програних матчів	Середня кількість забитих голів	Середня кількість пропущених голів
12	Борнмут	0.8120	11	6	0.5277778	1.0277778	1.916667
13	Вест Гем	0.8500	10	7	0.5277778	1.0555556	1.444444
14	Ноттінгем Форест	0.8365	8	10	0.5000000	1.0000000	1.861111
15	Евертон	0.8465	7	11	0.5000000	0.8888889	1.555556
16	Лідс	0.8460	7	10	0.5277778	1.2777778	1.972222
17	Лестер	0.8790	8	6	0.6111111	1.3611111	1.861111
18	Саутгемптон	0.8933	6	6	0.6666667	0.8611111	1.833333

Рис. 3.10. Поелементне входження команд до 1 групи

При виконанні другої строки ми отримаємо ті команди, які увійшли до групи команд-переможців (рис. 3.11):

Diplom.R | football[football_kmeans$cluster == 2,] | football[football_kmeans$cluster == 1,] | football_scale

Filter

	Назва команди	Відсоток доступних ключових гравців	Кількість перемог	Кількість нічиїх	Відсоток програних матчів	Середня кількість забитих голів	Середня кількість пропущених голів
1	Манчестер Сіті	0.960	28	4	0.1111111	2.555556	0.8611111
2	Арсенал	0.920	25	6	0.1388889	2.305556	1.1666667
3	Ньюкасл Юнайтед	0.950	19	12	0.1388889	2.027778	0.8611111
4	Ліверпуль	0.915	19	8	0.2500000	1.944444	1.1666667

Рис. 3.11. Поелементне входження команд до 2 групи

І, логічно, при запуску третьої строки коду ми отримаємо команди, які мають високий шанс зіграти внічию (рис. 3.12):

Diplom.R | football[football_kmeans$cluster == 3,] | football[football_kmeans$cluster == 2,] | football[f

Filter

	Назва команди	Відсоток доступних ключових гравців	Кількість перемог	Кількість нічиїх	Відсоток програних матчів	Середня кількість забитих голів	Середня кількість пропущених голів
5	Брайтон	0.9200	17	7	0.3333333	1.833333	1.250000
6	Тоттенгем	0.8960	17	6	0.3611111	1.805556	1.638889
7	Астон Вілла	0.8560	17	6	0.3611111	1.333333	1.222222
8	Брентфорд	0.9150	16	9	0.3055556	1.500000	1.250000
9	Фулгем	0.8560	15	6	0.4166667	1.444444	1.361111
10	Челсі	0.8546	11	10	0.4166667	1.000000	1.138889
11	Кристал Пелес	0.9150	11	10	0.4166667	1.027778	1.277778

Рис. 3.12. Поелементне входження команд до 3 групи

І останнє, що з вагомого ми ще можемо зробити у рамках цього аналізу, це з’ясувати різноманітні описові статистики для кластерного аналізу з допомогою функцій str(football_kmeans) та show(football_kmeans). Під час виконання першої функції ми можемо побачити які елементи до якого кластеру розміщені, бачимо центри кластерів, також ця функція показує які критерії ми використовували, дасть різноманітну інформацію стосовно розмірів відповдних кластерів і значення відповідних дисперсій також приведено (рис. 3.13):

```
> str(football_kmeans)
List of 9
 $ cluster     : int [1:18] 2 2 2 2 3 3 3 3 3 3 ...
 $ centers     : num [1:3, 1:6] -0.7793 1.2375 0.0722 -0.9288 1.3876 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "1" "2" "3"
  .. ..$ : chr [1:6] "Відсоток доступних ключових гравців" "Кількість перемог" "Кількість
нічиїх" "Відсоток програних матчів" ...
 $ totss       : num 102
 $ withinss    : num [1:3] 12.4 11.3 12.8
 $ tot.withinss: num 36.5
 $ betweenss   : num 65.5
 $ size        : int [1:3] 7 4 7
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

Рис. 3.13. Результат виконання функції str

Дивлячись результати останньої функції, ми можемо побачити знову ж таки розмір кластерів, з’ясувати середні значення по кластерам по відповідним факторам, знову таки знайти кластерний вектор і знайти відповідні велечини по диспресії (рис. 3.14):

```
> show(football_kmeans)
K-means clustering with 3 clusters of sizes 7, 4, 7

Cluster means:
  Відсоток доступних ключових гравців Кількість перемог Кількість нічиїх Відсоток програних матчів
1                          -0.77932720        -0.9288315       0.09816918
0.9688050
2                           1.23746587         1.3875836      -0.12271148
1.4566188
3                           0.07220385         0.1359266      -0.02804834
0.1364514
  Середня кількість забитих голів Середня кількість пропущених голів
1                       -0.76508528                          0.9981738
2                        1.46803166                         -1.1594508
3                       -0.07378995                         -0.3356305

Clustering vector:
 [1] 2 2 2 2 3 3 3 3 3 3 3 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 12.40549 11.26844 12.82591
 (between_SS / total_SS =  64.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

Рис. 3.14. Результат виконання функції show

Підсумовуючи, ми провели побудову кластерного аналізу за методом к-середніх та з’ясували розподіл початкових даних (тобто команд) по своїх кластерах, що знадобиться нам для подальшої побудови нових моделей.

3.2. Розробка дискримінантної моделі прогнозування результату футбольної гри та оцінка її якості у середовищі Statistica

Інформаційною базою дослідження є такі дані, як: відсоток доступних ключових гравців (тобто не враховуючи тих, хто травмований або дискваліфікований), кількість перемог, нічиїх і відосток поразок для кожної з команд, середня кількість забитих та пропущених голів для кожної з команд. Треба додати, що дані зібрані за 36 періодів (тобто за 36 зіграних матчів). Значення результуючої змінної задані наступним чином: W – перемога, D – нічия, L – поразка. Кожній команді дано відповідне значення результуючої змінної за допогою раніше побудованої моделі кластерного аналізу. Дані взяті за результатами «Англійської Премьєр Ліги сезону 2022-2023». Вхідні дані наведені на рис. 3.15.

	1 Відсоток доступних ключових гравців	2 Кількість перемог	3 Кількість нічиїх	4 Відсоток програних матчів	5 Середня кількість забитих голів	6 Середня кількість пропущени х голів	7 Клас
Манчестер Сіті	0,96	28	4	0,11111111	2,55555556	0,86111111	W
Арсенал	0,92	25	6	0,13888889	2,30555556	1,16666667	W
Ньюкасл Юнайтед	0,95	19	12	0,13888889	2,02777778	0,86111111	W
Ліверпуль	0,915	19	8	0,25	1,94444444	1,16666667	W
Брайтон	0,92	17	7	0,33333333	1,83333333	1,25	D
Тоттенгем	0,896	17	6	0,36111111	1,80555556	1,63888889	D
Астон Вілла	0,856	17	6	0,36111111	1,33333333	1,22222222	D
Брентфорд	0,915	13	14	0,30555556	1,5	1,25	D
Фулгем	0,856	15	6	0,41666667	1,44444444	1,36111111	D
Челсі	0,8546	11	10	0,41666667	1	1,13888889	D
Крістал Пелес	0,915	11	10	0,41666667	1,02777778	1,27777778	D
Борнмут	0,812	11	6	0,52777778	1,02777778	1,91666667	L
Вест Гем	0,85	10	7	0,52777778	1,05555556	1,44444444	L
Ноттінгем Форест	0,8365	8	10	0,5	1	1,86111111	L
Евертон	0,8465	7	11	0,5	0,88888889	1,55555556	L
Лідс	0,846	7	10	0,52777778	1,27777778	1,97222222	L
Лестер	0,879	8	6	0,61111111	1,36111111	1,86111111	L
Саутгемтон	0,8933	6	6	0,66666667	0,86111111	1,83333333	L

Рис. 3.15. Вхідні дані

Для початку шукаємо дискримінантний аналіз у середовищі Statistica (рис. 3.16–3.17):

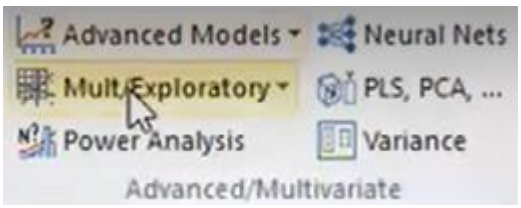

Рис. 3.16. Список варіантів аналізу

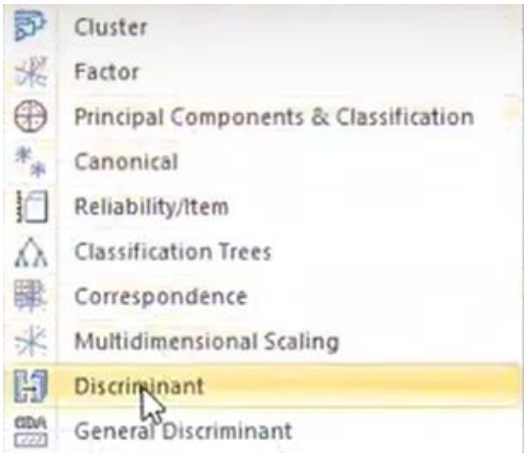

Рис. 3.17. Дискримінантний аналіз в Statistica

Далі обираємо незалежні та групуючи значення за допомогою кнопки «Variables» (рис. 3.18). Незалежні значення – статистичні значення за матч. В якості змінної для групування – величина класу (рис. 3.19):

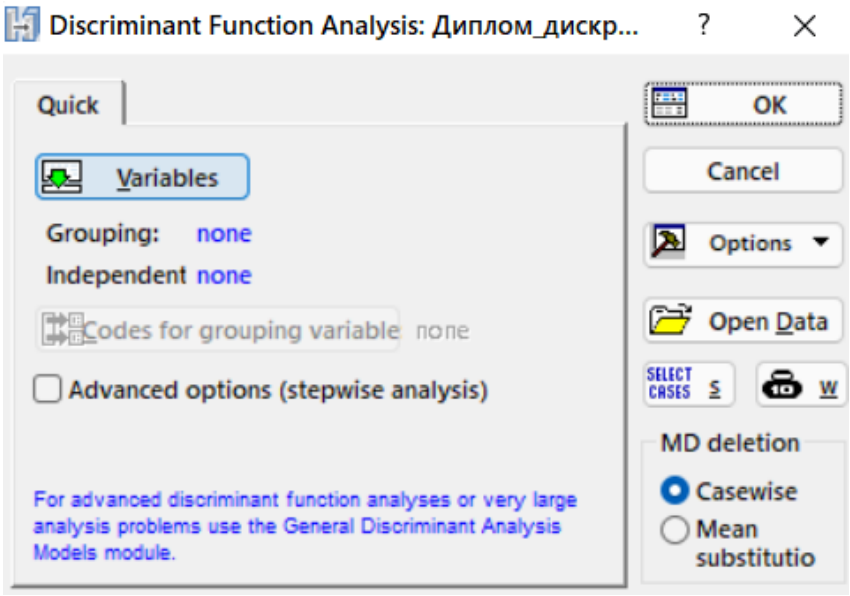

Рис. 3.18. Вибір значень

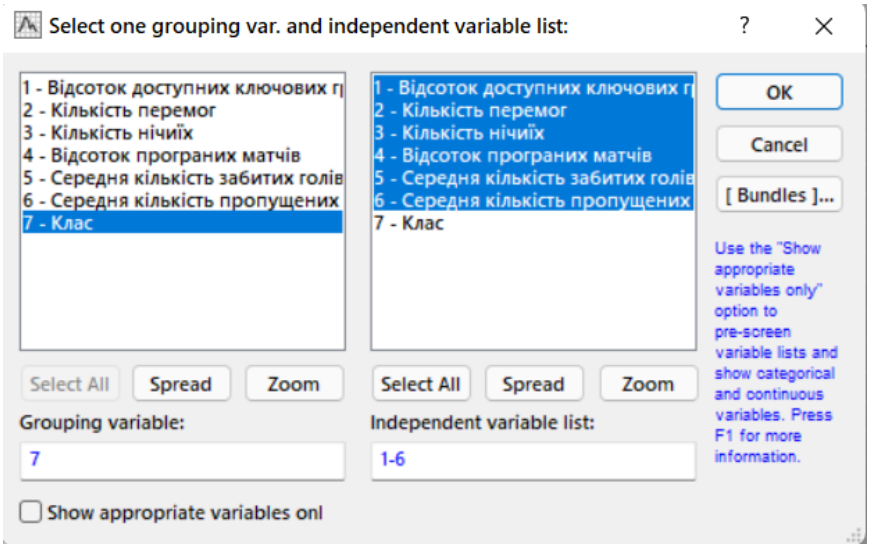

Рис. 3.19. Розподіл між значеннями

Після вибору значень натискаємо ОК. І запускаємо сам, безпосередньо, дискримінантний аналіз, який надасть нам перші результати (рис. 3.20). Тут можна побачити коефіцієнт Лямбди Уілкса, який характеризує якість дискримінації та змінюється в межах від 0 до 1. Якщо значення ближче до 0 – то воно визначає хорошу якість дискримінації. Також бачимо значення критерія Фішера, яке також використовується для оцінки адекватності моделі. За нашими результатами можемо сказати про доволі хороше дискримінаційне розбиття. Отже наша модель є статистично значущою.

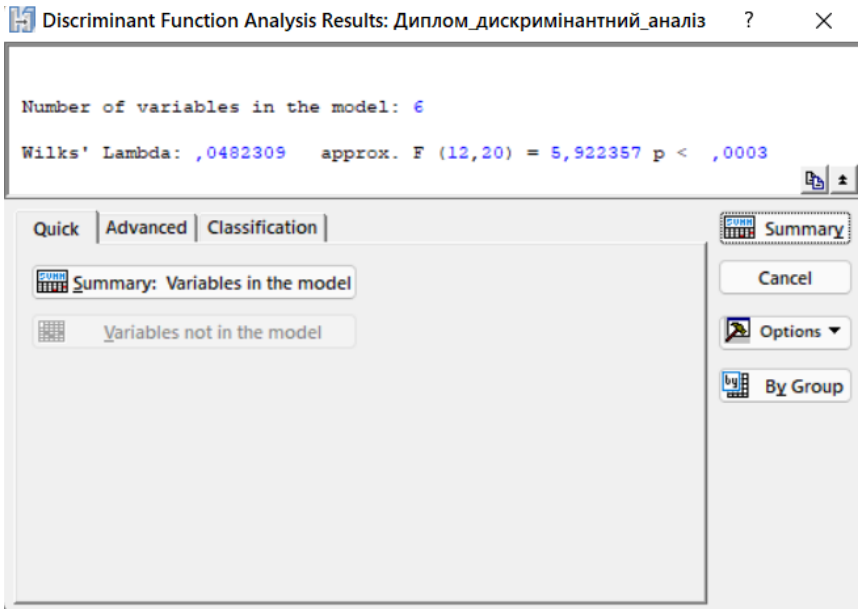

Рис. 3.20. Критерії Лямбди Уілкса та Фішера

Далі робимо пошук дискримінантних функцій, обравши вкладку Classification і натиснувши кнопку Classification function (рис. 3.21):

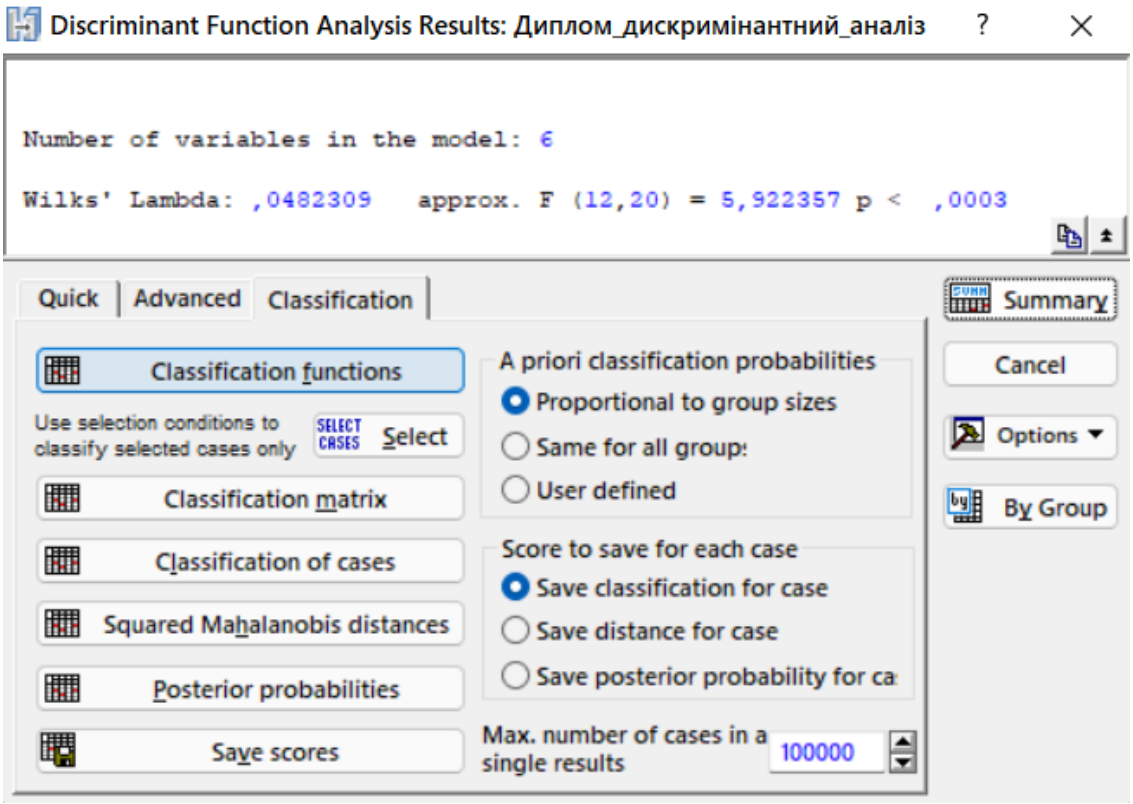

Рис. 3.21. Пошук дискримінантних функцій

Тобто, це як раз ті самі рівняння, які відповідають за кожну із груп. Наприклад, якщо розглядати по класу W, то він матиме таке рівняння (-4603,79 + + 1100,25X1 +244,94X2+213,96X3+8591,73X4-216,20X5+101,18X6). Рівняння за класом D має такий вигляд (-4755,45 + 1011,69X1 + 250,78X2 + + 219,85X3 +8880,59X4-218,69X5+105,54X6). Рівняння за класом L має такий вигляд (-4770,49 + 937,64X1 + 251,32X2 + + 221,12X3 + 8974,76X4-213,21X5+113,73X6), (рис. 3.22):

Variable	Classification Functions; grouping: Кла		
	L p=,38889	D p=,38889	W p=,22222
Відсоток доступних ключових гравців	937,64	1011,69	1100,25
Кількість перемог	251,32	250,78	244,94
Кількість нічиїх	221,12	219,85	213,96
Відсоток програних матчів	8974,76	8880,59	8591,73
Середня кількість забитих голів	-213,21	-218,69	-216,20
Середня кількість пропущених голів	113,73	105,54	101,18
Constant	-4770,49	-4755,45	-4603,79

Рис. 3.22. Дискримінантні функції

В рамках дискримінантного аналізу, також логічно знаходити так звану класифікаційну матрицю. Ця матриця дає інформацію про кількість і відсоток коректно класифікованих спостережень по кожній із груп (рис. 3.23):

Group	Classification Matrix (Spreadsheet1_(Recovered)) Rows: Observed classifications Columns: Predicted classifications			
	Percent Correct	L p=,38889	D p=,38889	W p=,22222
L	100,0000	7	0	0
D	100,0000	0	7	0
W	100,0000	0	0	4
Total	100,0000	7	7	4

Рис. 3.23. Класифікаційна матриця

Бачимо з результатів, що всі із визначених 18 об’єктів на 100% відповідають лінійним дискримінантним функціям.

Далі, можемо переглянути класифікацію спостережень за допомогою кнопки Classification of cases. Дане меню дає таблицю з класифікації для кожного спостереження, тобто воно показує в який клас і чому був доданий той чи інший елемент (рис. 3.24):

Case	Classification of Cases (Spreadsheet1_(Recovered)) Incorrect classifications are marked with *			
	Observed Classif.	1 p=,38889	2 p=,38889	3 p=,22222
Манчестер Сіті	W	W	D	L
Арсенал	W	W	D	L
Ньюкасл Юнайтед	W	W	D	L
Ліверпуль	W	W	D	L
Брайтон	D	D	W	L
Тоттенгем	D	D	L	W
Астон Вілла	D	D	L	W
Брентфорд	D	D	L	W
Фулгем	D	D	L	W
Челсі	D	D	L	W
Крістал Пелес	D	D	L	W
Борнмут	L	L	D	W
Вест Гем	L	L	D	W
Ноттінгем Форест	L	L	D	W
Евертон	L	L	D	W
Лідс	L	L	D	W
Лестер	L	L	D	W
Саутгемтон	L	L	D	W

Рис. 3.24. Таблиця класифікацій спостережень

І важливим є ще одне значення, яке може розраховуватися в Статистиці – це апостеріорні ймовірності (кнопка Posterior probabilities). Вони відповідають таблиці ймовірностей, яка показує приналежність кожного спостережень до відповідної групи. Тобто, з точки зору теорії ймовірностей показується, яка ймовірність потрапляння окремого спостереження до відповідної групи (рис.3.25).

Case	Posterior Probabilities (Spreadsheet1_(Recovered)) Incorrect classifications are marked with *			
	Observed Classif.	L p=,38889	D p=,38889	W p=,22222
Манчестер Сіті	W	0,000000	0,000000	1,000000
Арсенал	W	0,000000	0,000004	0,999996
Ньюкасл Юнайтед	W	0,000000	0,000000	1,000000
Ліверпуль	W	0,000000	0,106273	0,893727
Брайтон	D	0,000001	0,989701	0,010297
Тоттенгем	D	0,000667	0,999308	0,000025
Астон Вілла	D	0,000032	0,999967	0,000001
Брентфорд	D	0,000019	0,999981	0,000000
Фулгем	D	0,011510	0,988490	0,000000
Челсі	D	0,003287	0,996713	0,000000
Крістал Пелес	D	0,000137	0,999862	0,000001
Борнмут	L	0,999915	0,000085	0,000000
Вест Гем	L	0,972553	0,027447	0,000000
Ноттінгем Форест	L	0,999577	0,000423	0,000000
Евертон	L	0,990423	0,009577	0,000000
Лідс	L	0,999991	0,000009	0,000000
Лестер	L	0,999994	0,000006	0,000000
Саутгемтон	L	0,999995	0,000005	0,000000

Рис. 3.25. Таблиця апостеріорних ймовірностей

Таким чином, попередньо проведені дослідження дозволяють зробити висновок – побудована на наших даних модель є статистично значущою.

Отже, для того, щоб зрозуміти, куди віднесуться нові дві команди, які ми можемо додати в даний аналіз, потрібно спочатку їх додати у вхідні дані (рис.3.26):

	1 Відсоток доступних ключових гравців	2 Кількість перемог	3 Кількість нічиїх	4 Відсоток програних матчів	5 Середня кількість забитих голів	6 Середня кількість пропущених голів	7 Клас
Манчестер Сіті	0,96	28	4	0,11111111	2,55555556	0,86111111	W
Арсенал	0,92	25	6	0,13888889	2,30555556	1,16666667	W
Ньюкасл Юнайтед	0,95	19	12	0,13888889	2,02777778	0,86111111	W
Ліверпуль	0,915	19	8	0,25	1,94444444	1,16666667	W
Брайтон	0,92	17	7	0,33333333	1,83333333	1,25	D
Тоттенгем	0,896	17	6	0,36111111	1,80555556	1,63888889	D
Астон Вілла	0,856	17	6	0,36111111	1,33333333	1,22222222	D
Брентфорд	0,915	13	14	0,30555556	1,5	1,25	D
Фулгем	0,856	15	6	0,41666667	1,44444444	1,36111111	D
Челсі	0,8546	11	10	0,41666667	1	1,13888889	D
Крістал Пелес	0,915	11	10	0,41666667	1,02777778	1,27777778	D
Борнмут	0,812	11	6	0,52777778	1,02777778	1,91666667	L
Вест Гем	0,85	10	7	0,52777778	1,05555556	1,44444444	L
Ноттінгем Форест	0,8365	8	10	0,5	1	1,86111111	L
Евертон	0,8465	7	11	0,5	0,88888889	1,55555556	L
Лідс	0,846	7	10	0,52777778	1,27777778	1,97222222	L
Лестер	0,879	8	6	0,61111111	1,36111111	1,86111111	L
Саутгемтон	0,8933	6	6	0,66666667	0,86111111	1,83333333	L
Манчестер Юнайтед	0,9756	20	6	0,27777778	1,41666667	1,13888889	
Вулверхемптон	0,8456	11	7	0,5	0,83333333	1,44444444	

Рис. 3.26. Таблиця вхідних даних з двома доданими командами

Тепер повторюємо початкові дії, запускаючи дискримінантний аналіз, і через апостеріорні ймовірності дивимося, до яких класів і з якою ймовірністю віднесуться наші додані об’єкти (рис.3.27):

Case	Posterior Probabilities (Spreadsheet1_(Recovered)) Incorrect classifications are marked with *			
	Observed Classif.	L p=,38889	D p=,38889	W p=,22222
Манчестер Сіті	W	0,000000	0,000000	1,000000
Арсенал	W	0,000000	0,000004	0,999996
Ньюкасл Юнайтед	W	0,000000	0,000000	1,000000
Ліверпуль	W	0,000000	0,106273	0,893727
Брайтон	D	0,000001	0,989701	0,010297
Тоттенгем	D	0,000667	0,999308	0,000025
Астон Вілла	D	0,000032	0,999967	0,000001
Брентфорд	D	0,000019	0,999981	0,000000
Фулгем	D	0,011510	0,988490	0,000000
Челсі	D	0,003287	0,996713	0,000000
Крістал Пелес	D	0,000137	0,999862	0,000001
Борнмут	L	0,999915	0,000085	0,000000
Вест Гем	L	0,972553	0,027447	0,000000
Ноттінгем Форест	L	0,999577	0,000423	0,000000
Евертон	L	0,990423	0,009577	0,000000
Лідс	L	0,999991	0,000009	0,000000
Лестер	L	0,999994	0,000006	0,000000
Саутгемтон	L	0,999995	0,000005	0,000000
Манчестер Юнайтед	---	0,000000	0,014333	0,985667
Вулверхемптон	---	0,645401	0,354599	0,000000

Рис. 3.27. Апостеріорні ймовірності для команд Манчестер Юнайтед і Вулверхемптон

За результатами аналізу, бачимо що ймовірність Манчестер Юнайтед перемогти становить понад 98%. Ймовірність того, що Вулверхемптон програють, дивлячись на їхні показники, складає понад 65%. Вони мають шанси поборотися за нічию, але все ж ці шанси нижче за поразку – 35%. Отже можемо впевнено сказати – що Манчестер Юнайтед має перемагати у цьому протистоянні. Порівнюючи результати моделювання з реальними результатами (рис. 3.28), дійсно бачимо, що прогноз, який отриманий за моделлю, є вірним.

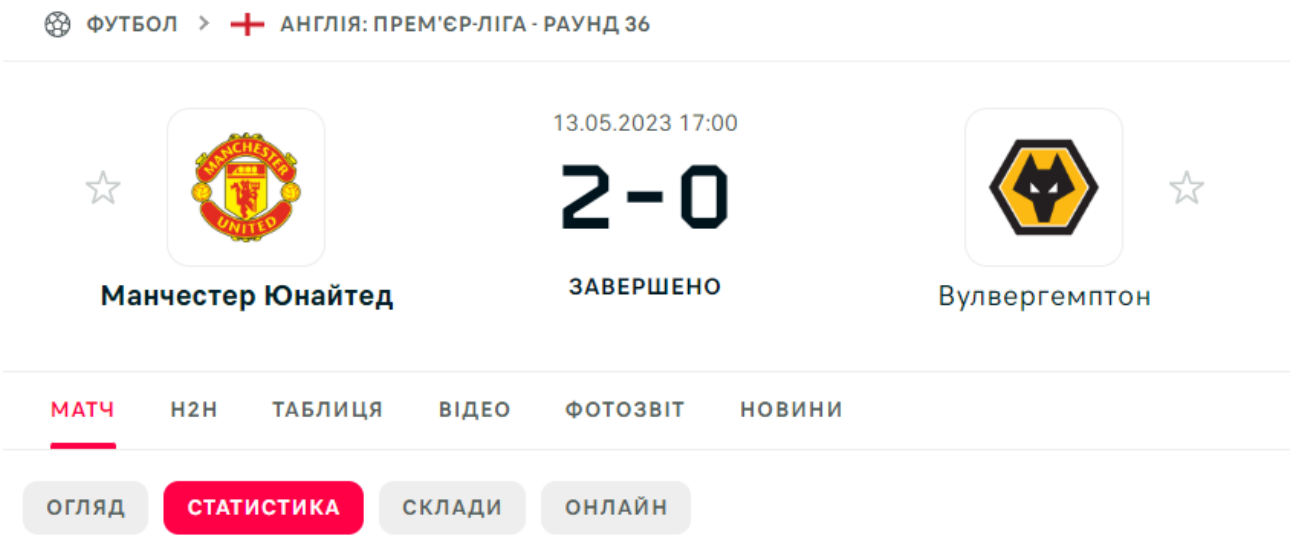

Рис. 3.28. Результат зіграного матчу

Виконавши аналіз моделі та проаналізувавши зібрані статистичні дані, я можу припустити, що Вулверхемптон програв за декількох причин. По переше, це відстутність декількох важливих гравців, які могли б допогти їм у цій грі. По друге, Вулверхемптон вкрай мало забиває голів. За цим показником вони одні з найгірших у лізі, у той самий час, коли оборона Манчестер Юнайтед одна з найкращих. І дійсно, ми бачимо що Вулверхемптон був навіть не в змозі забити бодай одного м’яча, що й призвело їх до поразки.

Отже, далі я пропоную розробити модель когнітивного моделювання, яка допоможе командам виявляти слабкі місця, як свої, так і свого супротивника, та підвищити якість прийняття рішень гравцями та тренерами, а також оптимізувати стратегії гри.

3.3. Застосування когнітивного моделювання для аналізу та поліпшення якості гри футбольної команди

Як нам вже відомо, спорт є однією з найбільш популярних діяльностей у світі, і футбол не є винятком. Цей вид спорту постійно розвивається, а разом з ним з'являються нові технології та інновації, які можуть покращити якість гри команд. Усвідомлення необхідності постійного поліпшення якості гри футбольних команд виникає з різних аспектів і вимагає уваги.

По-перше, поліпшена якість гри команди сприяє збільшенню її шансів на перемогу в матчах. Це може включати покращення тактичного розуміння, ефективного використання стратегій, розробку спеціальних групових тактик або підвищення навичок гравців.

По-друге, з поліпшенням якості гри команди гравці стають більш точними, швидкими та краще знають свої обов'язки на полі. Це допомагає уникати помилок, покращує передачу м'яча, зменшує втрати та збільшує кількість успішних атак і голів.

Також, команда, яка проявляє високу якість гри, здатна привернути більше вболівальників і збільшити зацікавленість спонсорів. Висока якість гри може приносити більше прибутку, популярності та рекламних можливостей для команди.

Розробка когнітивної моделі для поліпшення якості гри може допомогти з вирішенням таких питань:

1. Глибше розуміння ситуації: когнітивна модель дозволяє розкрити детальний характер впливів між чинниками, динаміку змін впливів та тимчасові зміни самого середовища гри. Це дозволяє команді отримати глибше розуміння процесу та факторів, що впливають на якість гри.

2. Прогнозування та планування: когнітивна модель надає можливість прогнозувати вплив різних факторів на якість гри та планувати стратегії та тактики для поліпшення гри. Вона дозволяє команді зробити осмислені рішення на основі аналізу впливу різних сценаріїв.

3. Точність і обґрунтованість: когнітивна модель базується на функціональних залежностях між чинниками, що дозволяє забезпечити точність та обґрунтованість аналізу. Вона надає об'єктивну основу для прийняття рішень та визначення оптимальних шляхів поліпшення якості гри.

4. Можливість впровадження інновацій: когнітивна модель дозволяє оцінювати вплив нових технологій та інновацій на якість гри. Вона створює можливість тестування різних сценаріїв та експериментів, що допомагає команді впроваджувати нові ідеї та знаходити інноваційні рішення.

5. Підвищення тренувального процесу: когнітивна модель може бути використана для покращення тренувального процесу команди. Вона дозволяє аналізувати та оцінювати ефективність тренувань, ідентифікувати слабкі місця та вдосконалювати підходи до тренувань для досягнення кращих результатів.

Починаючи побудову, спочатку маємо сформулювати матрицю взаємного впливу. Вона складається з таких факторів, як: Травми та дискваліфікації гравців, Форма команди (тобто кількість Перемог, Нічиїх і Поразок), Кількість забитих та пропущених м’ячів, Умови гри, Жовті та червоні картки. У ролі цільового чинника – рівень якості гри. Вхідні дані наведені на рис. 3.29:

	A	B	C	D	E	F	G	H	I
1	**Матриця впливу факторів**		(А)	(Б)	(В)	(Г)	(Д)	(Е)	(Є)
2	Рівень якості гри	(А)	0	0,7	0,5	-0,5	0,2	0	-0,4
3	Форма команди	(Б)	0,7	0	0	0	0	0	-0,2
4	Забиті м'ячі	(В)	0,5	0,6	0	0	0,2	-0,1	-0,2
5	Пропущені м'ячі	(Г)	0	0	0,3	0	-0,1	0,3	0,2
6	Умови гри	(Д)	0	0	0	0	0	0,05	0
7	Жовті та червоні картки	(Е)	0	0	0	0,1	0	0	0,05
8	Травми та дискваліфікації	(Є)	0	0	0	0	0	0,6	0

Рис. 3.29. Матриця впливу факторів

Починаємо розгляд процесу побудови когнітивної моделі для оцінки якості гри футбольної команди. З метою поліпшення команди, ми встановлюємо конкретні цілі. Наприклад, ми хочемо підвищити критерій форми команди на 10% і забиті м'ячі на 20%. Це означає, що ми можемо змінити нашу тактику та схему гри, додавши третього нападника до двох вже існуючих.

Проте, є й негативні аспекти, такі як травми та дискваліфікації. Для прикладу, уявімо, що один з наших ключових гравців травмований. Тому ми також підвищуємо на 10% показник травм та дискваліфікацій.

Далі, для кожного з цих показників ми формуємо покрокову таблицю, яка відображає зміни у відсотках (рис. 3.30). У цій таблиці ми наводимо формули, які нам допоможуть реалізувати встановлені цілі. Ця таблиця є важливим інструментом для розуміння і визначення наших кроків на шляху до поліпшення якості гри команди.

Всі ці дії спрямовані на розробку когнітивної моделі, яка дозволить нам краще розуміти вплив різних факторів на якість гри команди і виявити оптимальні стратегії для досягнення бажаних результатів.

10	**Зміна у % покроково**	0	1
11	Рівень якості гри	0	=МУМНОЖ($C$2:$I$8;B11:B17)
12	Форма команди	10	=МУМНОЖ($C$2:$I$8;B11:B17)
13	Забиті м'ячі	20	=МУМНОЖ($C$2:$I$8;B11:B17)
14	Пропущені м'ячі	0	=МУМНОЖ($C$2:$I$8;B11:B17)
15	Умови гри	0	=МУМНОЖ($C$2:$I$8;B11:B17)
16	Жовті та червоні картки	0	=МУМНОЖ($C$2:$I$8;B11:B17)
17	Травми та дискваліфікації	10	=МУМНОЖ($C$2:$I$8;B11:B17)

Рис. 3.30. Формули розрахунку покрокової зміни у відсотках

Тепер перейдемо до оцінки результатів реалізації цих формул, звертаючи увагу на важливу особливість - їх необхідно застосовувати до тих пір, поки значення не наблизяться до нуля або до моменту монотонного зростання або спадання (рис. 3.31).

Результати реалізації цих формул є важливими в нашому аналізі, оскільки вони надають нам об'єктивну інформацію про ефективність наших дій та дозволяють налаштовувати наші стратегії відповідно до змінюючихся умов та потреб команди. Таким чином, ми використовуємо ці формули як інструмент для об'єктивного вимірювання та оцінки прогресу, який ми досягаємо у поліпшенні якості гри команди.

10	**Зміна у % покроково**	0	1	2	3	4	5	6	7	8	9	10	11	12
11	Рівень якості гри	0	13	-3,4	8,205	-1,0158	4,57425	0,44196	2,64369	0,94412	1,67021	0,98365	1,1548	0,85992
12	Форма команди	10	-2	9,1	-2,44	5,6475	-0,729	3,17663	0,29414	1,83973	0,64945	1,16179	0,68017	0,80246
13	Забиті м'ячі	20	4	5,25	3,625	2,5355	2,843	1,81378	2,10394	1,47884	1,55928	1,21104	1,1763	0,97493
14	Пропущені м'ячі	0	8	1,35	1,8725	1,2245	0,84128	0,91528	0,58587	0,67019	0,47301	0,4958	0,38609	0,37404
15	Умови гри	0	0	0,025	0,04	0,0075	0,01056	0,00635	0,00452	0,00477	0,00307	0,00349	0,00246	0,00258
16	Жовті та червоні картки	0	0,5	0,8	0,15	0,21125	0,12695	0,09047	0,09534	0,0613	0,06988	0,04914	0,05168	0,04008
17	Травми та дискваліфікації	10	0	0,3	0,48	0,09	0,12675	0,07617	0,05428	0,0572	0,03678	0,04193	0,02948	0,03101

Рис. 3.31. Результат виконання попередніх формул

Для створення графіка, що відображатиме динаміку зміни значень факторів, що впливають на рівень якості гри команди, спочатку складемо таблицю даних, що відображатиме зміни у відсотках зростаючим підсумком. Це можна зробити за допомогою формули (B20+100)*(1+C11/100)-100. Результатом буде така таблиця (рис. 3.32):

19	**Зміна у % зростаючим підсумком**	0	1	2	3	4	5	6	7	8	9	10	11	12
20	Рівень якості гри	0	13	9,158	18,1144	16,9147	22,2626	22,803	26,0495	27,2396	29,3647	30,6372	32,1458	33,2822
21	Форма команди	10	7,8	17,6098	14,7401	21,2201	20,3363	24,159	24,5242	26,8151	27,6387	29,1216	29,9998	31,043
22	Забиті м'ячі	20	24,8	31,352	36,1135	39,5647	43,5325	46,1359	49,2105	51,4171	53,7781	55,6404	57,4712	59,0064
23	Пропущені м'ячі	0	8	9,458	11,5076	12,873	13,8226	14,8644	15,5373	16,3117	16,8618	17,4412	17,8946	18,3356
24	Умови гри	0	0	0,025	0,06501	0,07251	0,08309	0,08944	0,09397	0,09874	0,1018	0,1053	0,10776	0,11035
25	Жовті та червоні картки	0	0,5	1,304	1,45596	1,67028	1,79935	1,89144	1,98858	2,0511	2,12242	2,1726	2,2254	2,26637
26	Травми та дискваліфікації	10	10	10,33	10,8596	10,9594	11,1	11,1846	11,245	11,3086	11,3495	11,3962	11,4291	11,4636

Рис. 3.32. Таблиця зміни значень у відсотках зростаючим підсумком

Ця таблиця надасть нам інформацію про динаміку змін кожного фактора відносно початкового значення. Вона допоможе нам візуалізувати та аналізувати зміни впливу цих факторів на якість гри команди з часом. Графік, побудований на основі цієї таблиці, дозволить нам легше спостерігати та аналізувати тенденції та виявляти зв'язки між факторами та рівнем якості гри.

Таким чином, ця таблиця є важливим інструментом у процесі моніторингу та аналізу динаміки зміни факторів, що впливають на якість гри футбольної команди. Вона надає нам конкретні числові дані, які можна використовувати для подальшого вдосконалення та прийняття рішень щодо поліпшення гри команди.

Тепер ми готові створити графік, який демонструє зміну значень факторів, що впливають на рівень якості гри (рис. 3.33):

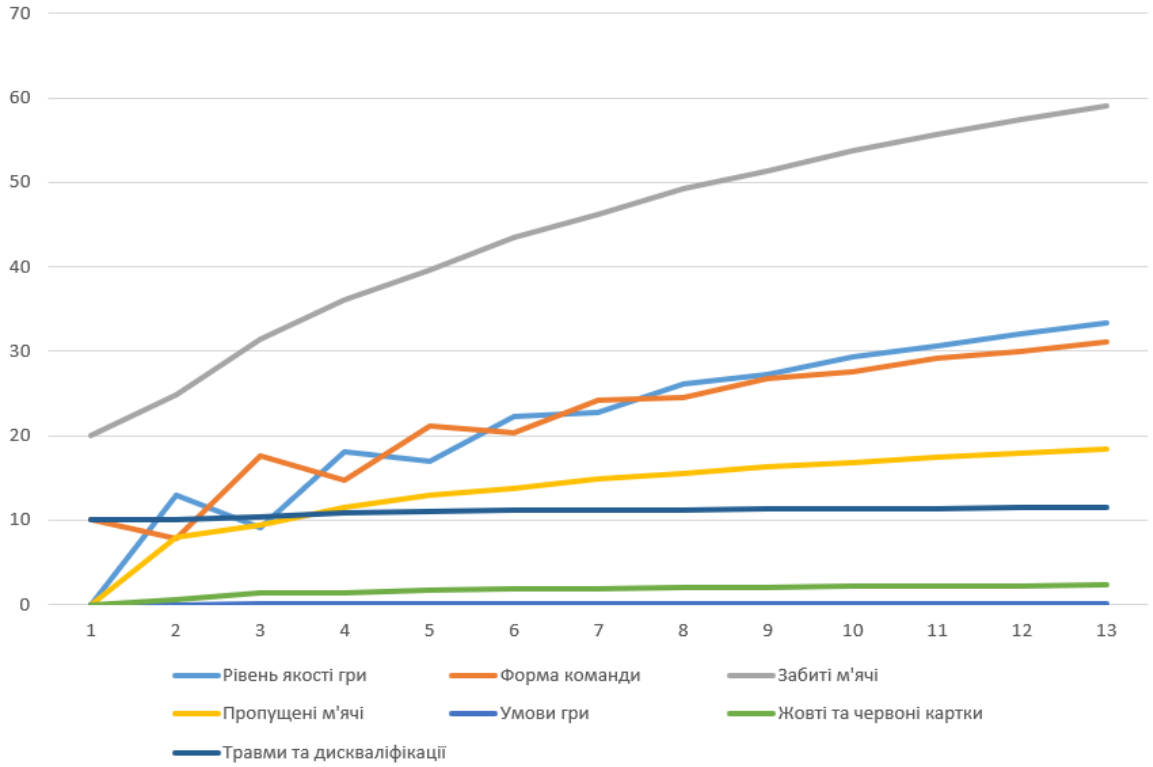

Рис. 3.33. Динаміка зміни значень факторів

Як видно з графіка, подібного до зображеного на рис. 3.33, з плином часу відбувається поступове зростання значення критерія Забитих м’ячів, що впливає на покращення якості гри команди. Це зростання, в свою чергу, має вплив на критерій Форми команди та критерій Пропущених м’ячів, які також показують підвищення протягом часу. Помірне збільшення значень критеріїв "Жовті і червоні картки" та "Травми і дискваліфікації" протягом тривалого періоду відповідає реальності. Всі ці зміни сприяють підвищенню рівня якості гри, навіть якщо початкове значення цього рівня було невисоким.

І, нарешті, проводимо розрахунок консонансу для оцінки якості отриманих результатів, рис. 3.34:

V	W	X	Y	Z
Фактор	Код	$\lvert\sum_k W_{mk} \cdot X_k(t)\rvert$	$\sum_k \lvert W_{mk} \cdot X_k(t)\rvert$	$C_m(t)$
Рівень якості гри	(А)	30,06185033	38,89335033	0,77293
Форма команди	(Б)	18,18283231	28,52088231	0,637527
Забиті м'ячі	(В)	28,57161043	28,57161043	1
Пропущені м'ячі	(Г)	17,18855022	17,18855022	1
Умови гри	(Д)	0,110299856	0,110299856	1
Жовті та червоні картки	(Е)	2,246080321	2,246080321	1
Травми та дискваліфікації	(Є)	1,323598268	1,323598268	1

Рис. 3.34. Розрахунок консонансу

Аналізуючи рис. 3.34, можна зробити висновок, що отримані результати є достатньо надійними і відповідають очікуваній інформації. Це підтверджується високими значеннями консонансу, які були розраховані. Такий висновок дозволяє нам мати довіру до отриманих даних і їх відповідності поставленим цілям.

Згідно з отриманими результатами дослідження, можна зробити висновок, що когнітивне моделювання є цінним інструментом для аналізу та покращення якості гри футбольної команди. Воно надає тренерам та гравцям глибше розуміння гри та можливість приймати обґрунтовані рішення, що сприяє підвищенню якості гри та досягненню кращих результатів. Наприклад, такі системи можуть визначати оптимальне розташування гравців на полі, передбачати траєкторію руху м'яча та прогнозувати наслідки різних стратегій гри.

Застосування когнітивного моделювання має потенціал допомогти тренерам та гравцям поліпшити якість гри футбольної команди і сприяти їх професійному розвитку. Ця технологія відкриває можливості досягнення кращих результатів і прокладає шлях до успіху. Це допомагає гравцям удосконалювати свої навички, підвищувати свій рівень гри і досягати переваги над суперниками. В цілому, використання когнітивного моделювання в футболі може мати значний вплив на підвищення якості гри команди і створити сприятливі умови для досягнення перемоги.

ВИСНОВКИ

У даній роботі було розглянуто методи прогнозування результатів футбольних матчів та оцінювання якості гри команд. Були реалізовані моделі за допомогою обраних методів, що дозволили вирішити завдання, поставлені у рамках роботи.

Для проведення дослідження взято статистичні показники, які характеризують результативність і рівень гри кожної з обраних команд, а саме: травми та дискваліфікації гравців, форма команди (тобто кількість перемог, нічиїх і поразок), кількість забитих та пропущених голів, умови гри, жовті та червоні картки.

За допомогою методів кластерного аналізу вдалося визначити оптимальну кількість кластерів для поділу початкових даних, тобто команд «Англійської Премьєр Ліги 2022-2023», побудувати дендограму дослідження та реалізувати метод к-середніх, що дозволило розділити команди на 3 групи, і в підсумку зробити оцінку якості кластерного аналізу. Тобто, кластерний аналіз допоміг зрозуміти, як краще розбити команди на 3 групи (переможці, нічийні та програвші) за тими статистичними характеристиками, які відображають рівень якості гри цих команд.

Дискримінантний аналіз дав змогу спробувати розробити модель прогнозування результату гри між двома обраними командами з цього ж чемпіонату. Власне, за результатами побудованої та проаналізованої моделі, можемо впевнено сказати що завдяки цьому методу багатовимірного аналізу можливо вдало будувати моделі прогнозування футбольних івентів.

В свою чергу, методи когнітивного моделювання допомогли побудувати модель аналізу та оптимізації якості гри футбольних команд. Побудована модель може бути корисна при реалізації інтелектуальних систем аналізу та управління грою, які дозволять швидко реагувати на зміни в грі та приймати оптимальні рішення. І це, безумовно, зацікавить багатьох людей, пов’язаних з футбольною сферою, бо поліпшення якості гри футбольних команд є дуже актуальною темою

в сучасному світі. Воно може мати вплив на різні аспекти спорту та бути корисним як для самого футболу, так і для спортсменів, глядачів та інвесторів.

Отже, в результаті виконання роботи, можна впевнено сказати, що на основі методів кластерного та дискримінантного аналізів була вдало побудована модель для прогнозування футбольного матчу, а також, завдяки когнітивному моделюванню досить вдало розроблено модель оцінювання та оптимізації якості гри футбольних команд.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Капустін О.І. Метод прогнозування результатів тенісних матчів на основі аналізу природної мови – [Електронний ресурс]. – Режим доступу: https://uni-sport.edu.ua/sites/default/files/vseDocumenti/diss_podrigalo_o.o.pdf

2. Штовба С.Д. Прогнозування результатів футбольних матчів на основі нечітких правил – [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/280064640_PROGNOZIROVANIE_REZULTATOV_FUTBOLNYH_MATCEJ_NA_OSNOVE_NECETKIH_PRAVIL

3. Галіцин С.В., Зіганшин О.З., Дубровін К.М., Єршов М.Є., Ткаченко П.А. Використання нейронних мереж для прогнозування відвідуваності футбольних матчів – [Електронний ресурс]. – Режим доступу: http://eprints.mdpu.org.ua/id/eprint/2557/1/sharov_poluhin_214_217.pdf

4. Бізнес-аналітика багатовимірних процесів: навчальний посібник / Т. С. Клебанова, Л. С. Гур'янова, Л.О. Чаговець та ін. – Харків.: ХНЕУ ім. С. Кузнеця, 2018. 272 с. URL: http://www.repository.hneu.edu.ua/bitstream/123456789/22020/1/2018-Клебанова%2CГурьянова%2C%20Чаговец%20и%20др.pdf

5. Спортивна аналітика як наука: що таке прогнозування – [Електронний ресурс]. – Режим доступу: https://kp.ua/sport/696641-fakt-sportyvnaia-analytyka-kak-nauka-chto-takoe-prohnozyrovanye

6. Основи економічного прогнозування – [Електронний ресурс]. – Режим доступу: https://buklib.net/books/27070/

7. Дослідження та розробка методів прогнозування з використанням імовірнісних нейронних мереж – [Електронний ресурс]. – Режим доступу: https://openarchive.nure.ua/server/api/core/bitstreams/02be6987-9226-4edd-b910-83955e90a1fc/content

8. Цаконас А.Д., Дуніас Г.Д., Штовба С.Д. Прогнозування результатів футбольних матчів за допомогою машини опорних векторів // Вісник Житомирського інженернотехнологічного інститута. – 2003. – №1.- С. 181–186.

Режим доступу: https://library.ztu.edu.ua/e-copies/VISNUK/24_I/181.pdf

9. Інформаційна технологія аналізу результатів спортивних подій – [Електронний ресурс]. – Режим доступу: https://essuir.sumdu.edu.ua/bitstream-download/123456789/79513/1/Monko_bac_rab.pdf;jsessionid=6CE08DCDA6842FEDEDA6C5EAD4DD5E54

10. Математичні моделі прогнозування результатів футбольних матчів – [Електронний ресурс]. – Режим доступу: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/26569/semenuk.pdf?sequence=1&isAllowed=y

11. Зеленцов А., Лобановський В. Методологічні основи розробки моделей тренувальних занять – [Електронний ресур]. – Режим доступу: https://www.researchgate.net/publication/323365954_Modeluvanna_v_sporti_ak_metod_doslidzenna

12. Метод виваженої суми показників для прогнозування футбольних матчів – [Електронний ресурс]. – Режим доступу: http://chexov.com.ua/video/2010881719-analiz-futbolnyh-matchey/

13. Методи прогнозування футбольних подій – [Електронний доступ]. – Режим доступу: https://er.knutd.edu.ua/bitstream/123456789/1906/1/td_2016_N1_14.pdf

14. Футбольні результати онлайн – [Електронний ресурс]. – Режим доступу: https://www.flashscore.ua/?rd=flashscore.com.ua

15. Багатовимірний статистичний аналіз. Завдання класифікації об'єктів: кластерний аналіз, дискримінантний аналіз – [Електронний ресурс]. – Режим доступу: http://ni.biz.ua/9/9_15/9_155153_mnogomerniy-statisticheskiy-analiz-zadachi-klassifikatsii-ob-ektov-klasterniy-analiz-diskriminantniy-analiz.html

16. Суть дискримінантного аналізу – [Електронний ресурс]. – Режим доступу: https://financial.lnu.edu.ua/wp-content/uploads/2015/10/78.pdf

17. Сутність і завдання дискримінантного аналізу. Обмеження та проблеми використання методів дискримінантного аналізу – [Електронний ресурс]. – Режим доступу: http://ebooks.git-elt.hneu.edu.ua/babap/5-1-id5-1.html

18. Методи дискримінантного аналізу. Алгоритм лінійного дискримінантного аналізу Фішера для двох класів. Перевірка якості дискримінації – [Електронний ресурс]. – Режим доступу: http://ebooks.git-elt.hneu.edu.ua/babap/5-2-id5-2.html

19. Особливості застосування методів кластерного аналізу – [Електронний ресурс]. – Режим доступу: http://ebooks.git-elt.hneu.edu.ua/babap/3-1-id3-1.html

20. Термінологія кластерного аналізу – [Електронний ресурс]. – Режим доступу: http://ebooks.git-elt.hneu.edu.ua/babap/3-2-id3-2.html

21. Міри подібності – [Електронний ресурс]. – Режим доступу: http://ebooks.git-elt.hneu.edu.ua/babap/3-3-id3-3.html

22. ДЕЯКІ АСПЕКТИ ЗАСТОСУВАННЯ КОГНІТИВНОГО МОДЕЛЮВАННЯ В ДЕРЖАВНОМУ УПРАВЛІННІ – [Електронний ресурс]. – Режим доступу: http://www.дудом.nayka.com.ua/?op=1&z=922

23. Когнітивне моделювання складних ситуацій – [Електронний ресурс]. – Режим доступу: https://pns.hneu.edu.ua/pluginfile.php/771289/mod_resource/content/1/%D0%A2%D0%B5%D0%BC%D0%B0%202%20%D0%9A%D0%BE%D0%B3%D0%BD%D1%96%D1%82%D0%B8%D0%B2%D0%BD%D0%B5%20%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8E%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F%20%D1%81%D0%BA%D0%BB%D0%B0%D0%B4%D0%BD%D0%B8%D1%85%20%D1%81%D0%B8%D1%82%D1%83%D0%B0%D1%86%D1%96%D0%B9.pdf

24. Когнітивна модель складної ситуації – [Електронний ресурс]. – Режим доступу: https://pns.hneu.edu.ua/pluginfile.php/771326/mod_resource/content/1/%D0%90%D0%9F%D0%9C%20%D0%9B%D0%B0%D0%B1%D0%B0%202%20%D0%9A%D0%BE%D0%B3%D0%BD%D1%96%D1%82%D0%B8%D0%B2%D0%BD%D0%B8%D0%B9%20%D0%B0%D0%BD%D0%B0%D0%BB%D1%96%D0%B7.pdf

25. Дискримінантний аналіз в ППП Statistica – [Електронний ресурс]. –

Режим доступу: https://ru.essays.club/%D0%A2%D0%BE%D1%87%D0%BD%D1%8B%D0%B5-%D0%BD%D0%B0%D1%83%D0%BA%D0%B8/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0/%D0%94%D0%B8%D1%81%D0%BA%D1%80%D0%B8%D0%BC%D0%B8%D0%BD%D0%B0%D0%BD%D1%82%D0%BD%D1%8B%D0%B9-%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7-%D0%B2-%D0%9F%D0%9F%D0%9F-12944.html

26. Знайомство з мовою програмування R – [Електронний ресурс]. – Режим доступу: https://rstudio-pubs-static.s3.amazonaws.com/378130_5600736fb2734e01bf109c83e6d83676.html

27. Кластерний аналіз в R/Cluster analysis in R – [Електронний ресурс]. – Режим доступу: https://www.youtube.com/watch?v=4HsD9tpC1tU&list=PLbueqh_j3hcDqo8XX7_H_WWK1l5BZxrMj&index=7

28. Теоретичні і методологічні основи економічного прогнозування. Зміст, предмет, метод і завдання курсу – [Електронний ресурс]. – Режим доступу: https://buklib.net/books/27070/

29. Історичні аспекти використання багатовимірного статистичного аналізу. Методи багатовимірного статистичного аналізу – [Електронний ресурс]. – Режим доступу: http://ebooks.git-elt.hneu.edu.ua/babap/1-2-id1-2.html

30. Статистичне моделювання – [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%BD%D0%B5_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8E%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F

31. Лінійний дискримінантний аналіз – [Електронний ресурс]. – Режим доступу: http://www.tsatu.edu.ua/kn/wp-content/uploads/sites/16/3_lab.3_dyskrymynantnyj_analyz.pdf

32. Лінійний розділювальний аналіз – [Електронний ресурс]. – Режим

доступу: https://uk.wikipedia.org/wiki/%D0%9B%D1%96%D0%BD%D1%96%D0%B9%D0%BD%D0%B8%D0%B9_%D1%80%D0%BE%D0%B7%D0%B4%D1%96%D0%BB%D1%8E%D0%B2%D0%B0%D0%BB%D1%8C%D0%BD%D0%B8%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D1%96%D0%B7

33. Когнітивне моделювання складних систем – [Електронний ресурс]. – Режим доступу: https://stud.com.ua/25055/menedzhment/kognitivne_modelyuvannya_skladnih_sistem

34. Кластерний аналіз: введення – [Електронний ресурс]. – Режим доступу: https://rpubs.com/AllaT/clust1

35. Когнітивне моделювання в Excel – [Електронний ресурс]. – Режим доступу: https://ena.lpnu.ua:8443/server/api/core/bitstreams/31059425-33c2-4a53-95f3-a72d3de426c9/content

36. ОБРОБКА ІНФОРМАЦІЇ ТАБЛИЧНИМ ПРОЦЕСОРОМ – [Електронний ресурс]. – Режим доступу: http://www.tsatu.edu.ua/kn/wp-content/uploads/sites/16/lekcija-6.pdf

37. Робота з мовою R – [Електронний ресурс]. – Режим доступу: https://rstudio-pubs-static.s3.amazonaws.com/378105_599bcd2892bf46498a6371290149267d.html

38. Робота з пакетом psych – [Електронний ресурс]. – Режим доступу: https://rpubs.com/nvasilenok/482977

39. Extract and Visualize the Results of Multivariate Data Analyses – [Електронний ресурс]. – Режим доступу: https://cran.r-project.org/web/packages/factoextra/readme/README.html

40. Кластеризация – [Електронний ресурс]. – Режим доступу: https://agricolamz.github.io/DS_for_DH/%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F.html

41. Довідкова карта з Data Mining – [Електронний ресурс]. – Режим

доступу: http://csc.knu.ua/media/study/asp/mod_probl_inf_tech_sys_analysis_ivohin/lecture/lec3.pdf

42. Повний список зручних пакетів в R – [Електронний доступ]. – Режим доступу: https://rstudio-pubs-static.s3.amazonaws.com/378105_599bcd2892bf46498a6371290149267d.html

43. Основи програмування в R – [Електронний доступ]. – Режим доступу: https://rpubs.com/AllaT/rprog-ggplot2

44. Аналіз даних з R – [Електронний доступ]. – Режим доступу: https://sociology.knu.ua/sites/default/files/course/materials/r1.pdf

45. Наочна Статистика – [Електронний доступ]. – Режим доступу: https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf

46. Таблиця «Чемпіонату Англійської Премьєр-Ліги сезону 2022-2023» – [Електронний доступ]. – Режим доступу: https://www.flashscore.ua/match/G8hKUpsm/#/standings/table/overall

ДОДАТКИ

ДОДАТОК А

Повний код реалізації кластерного аналізу в середовищі RStudio

```
install.packages("psych")
install.packages("factoextra")
install.packages("cluster")
install.packages("NbClust")
install.packages("clValid")
install.packages("ggplot2")
install.packages("MASS")
install.packages("gplots")
#Завантаження пакетів для реалізації кластерного аналізу роботи програми
library(psych)
library(factoextra)
library(cluster)
library(NbClust)
library(clValid)
library(ggplot2)
library(MASS)
library(gplots)
#КЛАСТЕРНИЙ АНАЛІЗ#
#Початкова обробка даних
View(football) #перевірка введення інформації в середовище
football <- as.data.frame(football)
football_scale <- scale(football[-1], center=T, scale=T) #проведення стандартизації даних
View(football_scale)
#Побудова дендрограми
football_res <- hclust(dist(football_scale, method = "euclidean"), method = "ward.D2") #запускаємо алгоритм Уорда
```

```
graphs_vision <- cutree(football_res, k = 3) # отримали 3 візуальні кластери
plot(football_res, cex = 0.8) # побудова графіку
rect.hclust(football_res, k = 3, border = 2:4) # візуально підтверджується розбиття на 2 кластери, але так як нам треба розподілити команди на 3 групи, то робимо 3 кластери і додаємо рамки
#Перевірка оптимальної кількості кластерів
fviz_nbclust(football_scale, kmeans, method = "wss") #автоматична перевірка методом "ліктя"
fviz_nbclust(football_scale, kmeans, method = "silhouette") #автоматична перевірка силуетним методом
fviz_nbclust(football_scale, kmeans, method = "gap_stat") #автоматична перевірка методом gap-статистики
#Реалізація К-середніх
football_kmeans <- kmeans(football_scale, 3, nstart = 18)
fviz_cluster(football_kmeans, data = football_scale, frame.type = "convex")+theme_minimal()
#Поелементне входження в кластери
View(football[football_kmeans$cluster==1,])
View(football[football_kmeans$cluster==2,])
View(football[football_kmeans$cluster==3,])
#описова статистика кластерів
str(football_kmeans)
show(football_kmeans)
```