

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук
(повна назва)
Кафедра _____ Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____ перший (бакалаврський)

Розробка інтелектуальної системи прогнозування результативних дій
у спортивній аналітиці на основі методів машинного навчання
(тема)

Виконав:
здобувач _____ четвертого _____ року навчання,
групи _____ ІТШ-22-2

_____ Артем САБАЛАЄВ
(власне ім'я, прізвище)

Спеціальність 122 Комп'ютерні науки

(код і повна назва спеціальності)

Тип програми _____ освітньо-професійна

Освітня програма _____ Штучний інтелект

(повна назва освітньої програми)

Керівник _____ ас. Дмитро МАЛЄЄВ
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ

(підпис)

_____ Лариса ЧАЛА
(власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____

Кафедра _____ Штучного інтелекту _____

Рівень вищої освіти _____ перший (бакалаврський) _____

Спеціальність _____ 122 Комп'ютерні науки _____
(код і повна назва)

Тип програми _____ освітньо-професійна _____

Освітня програма _____ Штучний інтелект _____
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

« _____ » _____ 20 ____ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві _____ Сабалаєву Артему Руслановичу _____
(прізвище, ім'я, по батькові)

1. Тема роботи _____ Розробка інтелектуальної системи прогнозування результативних дій у спортивній аналітиці на основі методів машинного навчання _____

затверджена наказом університету від _____ 5 _____ червня _____ 20 26 р. № 488Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії _____ 8 _____ липня _____ 20 26 р.

3. Вихідні дані до роботи _____ дані Інтернет-джерел, відкритий набір футбольних event-даних StatsBomb Open Data, що містить інформацію про події матчів провідних футбольних турнірів; методи машинного навчання для задач бінарної класифікації; мова програмування Python; бібліотеки Pandas, NumPy, Scikit-learn, XGBoost, CatBoost, Matplotlib, Seaborn; середовище розробки Google Colab; фреймворк Streamlit для реалізації програмної системи. _____

4. Перелік питань, що потрібно опрацювати в роботі _____

1) Аналіз предметної галузі та постановка задачі дослідження _____

2) Методи та моделі прогнозування результативних дій у спортивній аналітиці _____

3) Розробка інтелектуальної системи прогнозування результативності ударів _____

4) Експериментальне дослідження та аналіз результатів _____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / термін виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	08.06.2026	виконано
2	Аналіз предметної галузі та постановка задачі	08.06.2026–10.06.2026	виконано
3	Теоретичне дослідження методів машинного навчання у спортивній аналітиці	10.06.2026–12.06.2026	виконано
4	Формування та підготовка набору даних футбольних подій	13.06.2026–15.06.2026	виконано
5	Розробка та навчання моделей прогнозування результативності ударів	15.06.2026–19.06.2026	виконано
6	Проведення експериментальних досліджень та оцінювання якості моделей	20.06.2026–22.06.2026	виконано
7	Розробка інтелектуальної системи та вебінтерфейсу на Streamlit	22.06.2026–24.06.2026	виконано
8	Оформлення пояснювальної записки	25.06.2026–27.06.2026	виконано
9	Нормоконтроль	28.06.2026–30.06.2026	виконано
10	Рецензування	01.07.2026	виконано
11	Попередній захист	04.07.2026	виконано
12	Захист перед ЕК	08.07.2026	

Дата видачі завдання 8 червня 20 26 р.

Здобувач АСР
(підпис)

Керівник роботи _____
(підпис)

ас. Дмитро МАЛЄЄВ
(посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 112 с., 82 рис., 1 дод., 34 джерела.

КЛАСИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ, ПРОГНОЗУВАННЯ РЕЗУЛЬТАТИВНИХ ДІЙ, РЕЗУЛЬТАТИВНІСТЬ УДАРІВ, СПОРТИВНА АНАЛІТИКА, ФУТБОЛЬНА АНАЛІТИКА, EVENT-ДАНІ, EXPECTED GOALS.

Об'єкт дослідження – процес аналізу та прогнозування результативних дій у футбольній аналітиці.

Предмет дослідження – методи та моделі машинного навчання для прогнозування результативності ударів у футболі на основі подій матчів.

Мета роботи – розробка інтелектуальної системи прогнозування результативності ударів у футбольній аналітиці на основі методів машинного навчання.

Методи дослідження – методи машинного навчання, методи класифікації, статистичний аналіз даних, методи обробки спортивних даних, методи оцінювання якості моделей прогнозування, засоби програмної реалізації на Python.

У роботі розглянуто задачу прогнозування результативних дій у спортивній аналітиці із застосуванням методів машинного навчання. На основі відкритих футбольних event-даних сформовано набір даних для прогнозування результативності ударів. Реалізовано та досліджено моделі Logistic Regression, Random Forest, XGBoost, CatBoost і багатопаровий перцептрон. Проведено експериментальне оцінювання моделей та виконано їх порівняльний аналіз. Розроблено інтелектуальну систему прогнозування результативності ударів із вебінтерфейсом на базі Streamlit, що забезпечує аналіз реальних ігрових ситуацій та моделювання нових атакуювальних епізодів.

ABSTRACT

Bachelor's thesis contains: 112 pp., 82 fig., 1 ann., 34 references.

CLASSIFICATION, EVENT DATA, EXPECTED GOALS, FOOTBALL ANALYTICS, MACHINE LEARNING, PREDICTION OF SCORING ACTIONS, SHOT EFFECTIVENESS, SPORTS ANALYTICS.

The object of research is the process of analyzing and predicting scoring actions in football analytics.

The subject of research is machine learning methods and models for predicting shot effectiveness in football based on match event data.

The purpose of the thesis is to develop an intelligent system for predicting shot effectiveness in football analytics using machine learning methods.

Research methods include machine learning methods, classification methods, statistical data analysis, sports data processing methods, predictive model evaluation methods, and Python-based software development tools.

The thesis addresses the problem of predicting scoring actions in sports analytics using machine learning methods. A dataset for shot effectiveness prediction was created based on open football event data. Logistic Regression, Random Forest, XGBoost, CatBoost, and Multi-Layer Perceptron models were implemented and investigated. Experimental evaluation and comparative analysis of the models were carried out. An intelligent system for predicting shot effectiveness with a Streamlit-based web interface was developed, providing analysis of real match situations and simulation of new attacking scenarios.

ЗМІСТ

Вступ.....	8
1 Аналіз предметної галузі та постановка задачі дослідження	10
1.1 Особливості аналізу результативних дій у футбольній аналітиці	10
1.2 Аналіз існуючих методів та систем спортивної аналітики у футболі	17
1.3 Постановка задачі дослідження.....	31
2 Методи та моделі прогнозування результативних дій у спортивній аналітиці	33
2.1 Класичні статистичні методи аналізу спортивних даних	33
2.2 Методи машинного навчання у задачах спортивної аналітики	38
2.3 Методи представлення та обробки футбольних даних	43
2.4 Методи оцінювання якості моделей прогнозування	48
3 Розробка інтелектуальної системи прогнозування результативності ударів	54
3.1 Формування та підготовка датасету футбольних ударів	55
3.1.1 Формування набору футбольних матчів та event-даних.....	55
3.1.2 Формування датасету ударів.....	57
3.1.3 Дослідницький аналіз даних	60
3.1.4 Попередня обробка даних для моделей машинного навчання	65
3.2 Реалізація моделей машинного навчання.....	67
3.2.1 Logistic Regression.....	67
3.2.2 Random Forest	68
3.2.3 XGBoost.....	69
3.2.4 CatBoost.....	71
3.2.5 MLP	73
3.3 Розробка користувацького інтерфейсу системи	74
4 Експериментальне дослідження та аналіз результатів.....	79
4.1 Результати навчання моделей машинного навчання	79
4.1.1 Logistic Regression.....	79

4.1.2 Random Forest	83
4.1.3 XGBoost.....	86
4.1.4 CatBoost.....	89
4.1.5 MLP	92
4.2 Порівняльний аналіз моделей прогнозування	96
4.3 Аналіз роботи системи	99
4.3.1 Аналіз прогнозування на реальних ігрових ситуаціях.....	99
4.3.2 Аналіз роботи системи при моделюванні ударів	103
Висновки	106
Перелік джерел посилання	108
Додаток А Відомість кваліфікаційної роботи	112

ВСТУП

Спортивна аналітика є одним із напрямів, у якому активно застосовуються сучасні інформаційні технології, методи інтелектуального аналізу даних та засоби штучного інтелекту. У професійному спорті дедалі більшу роль відіграє не лише фізична підготовка спортсменів або тактичні рішення тренерського штабу, а й здатність ефективно аналізувати великі обсяги статистичної інформації, що формується під час проведення матчів. Використання цифрових технологій дозволяє здійснювати детальний аналіз ігрових ситуацій, оцінювати ефективність командних та індивідуальних дій, виявляти закономірності у поведінці гравців і формувати аналітичні висновки на основі реальних даних.

Особливого поширення спортивна аналітика набула у футболі, де кожен матч супроводжується значною кількістю подій та статистичних показників. Сучасні системи збору даних дозволяють фіксувати координати ігрових дій, передачі, удари, володіння м'ячем, характеристики атак, інтенсивність гри та інші параметри, що описують перебіг матчу. Аналіз подібної інформації створює можливість більш глибокого дослідження ігрових процесів, оцінювання ефективності атакувальних дій та виявлення факторів, що впливають на результативність команди.

Однією з актуальних задач спортивної аналітики є прогнозування результативних дій. Такі задачі безпосередньо пов'язані з оцінюванням ймовірності успішного завершення атакувальних епізодів та аналізом факторів, що впливають на створення небезпечних моментів. Результати подібного аналізу можуть використовуватись для оцінювання ефективності гравців, аналізу тактичних схем, підготовки до матчів та підтримки процесу прийняття рішень у спортивній діяльності. При цьому складність задачі прогнозування полягає у необхідності врахування великої кількості взаємопов'язаних параметрів, які можуть по-різному впливати на результат залежно від конкретної ігрової ситуації.

Традиційні статистичні методи аналізу спортивних даних мають низку обмежень під час роботи зі складними багатовимірними наборами даних. Використання лише базових статистичних показників не завжди дозволяє врахувати приховані залежності між параметрами гри або особливості розвитку атаквальних дій. У зв'язку з цим у спортивній аналітиці все активніше застосовуються методи машинного навчання, здатні автоматично виявляти закономірності у даних та формувати моделі прогнозування на основі великої кількості історичних спостережень.

Методи машинного навчання дозволяють здійснювати обробку великих обсягів спортивних даних, аналізувати характеристики ігрових ситуацій та прогнозувати ймовірність результативного завершення певних дій. Використання таких підходів забезпечує можливість підвищення точності аналізу, автоматизації процесів оцінювання та створення інтелектуальних систем підтримки прийняття рішень. Особливу увагу при цьому приділяють задачам класифікації та прогнозування на основі подій матчів, оскільки саме event-дані дозволяють детально описувати перебіг гри та взаємозв'язки між окремими ігровими епізодами.

Актуальність теми зумовлена зростанням ролі інтелектуального аналізу даних у професійному спорті, необхідністю підвищення точності прогнозування результативних дій та потребою у створенні програмних засобів, здатних автоматизувати обробку й аналіз спортивної інформації. Розробка інтелектуальної системи прогнозування результативних дій на основі методів машинного навчання дозволяє поєднати сучасні підходи до аналізу спортивних даних із практичними задачами футбольної аналітики та створити програмне рішення, придатне для дослідження й оцінювання ефективності атаквальних дій у футболі.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Особливості аналізу результативних дій у футбольній аналітиці

У сучасному професійному спорті аналітика стала одним із ключових інструментів підтримки процесу прийняття рішень. Використання статистичних даних, цифрових технологій та методів обробки інформації дозволяє більш об'єктивно оцінювати ефективність командних і індивідуальних дій, виявляти закономірності у поведінці гравців та аналізувати ігрові сценарії. Особливо помітним розвиток спортивної аналітики став у футболі, де велика кількість подій під час матчу формує значний обсяг даних для подальшого дослідження.

Традиційно оцінювання якості гри значною мірою базувалося на експертній думці тренерів, скаутів або спортивних аналітиків. Проте сучасний футбол дедалі більше орієнтується на data-driven підхід, у межах якого рішення приймаються не лише на основі суб'єктивних спостережень, а й з урахуванням статистичних показників, аналітичних моделей та результатів автоматизованого аналізу даних. Це стосується як оцінювання окремих гравців, так і аналізу тактичних схем, ефективності атакуювальних дій та прогнозування результатів ігрових епізодів [1].

Розвиток цифрових технологій та автоматизованих систем збору спортивної статистики суттєво змінив підходи до аналізу футбольних матчів. Якщо раніше основна частина аналітичної інформації формувалася вручну на основі перегляду матчів та експертних спостережень, то сучасні технології дозволяють фіксувати практично кожну дію, що відбувається на полі. Використання спеціалізованих систем відеотрекінгу, комп'ютерного бачення та сенсорних технологій забезпечує отримання великого обсягу деталізованих даних про перебіг гри. У результаті футбольний матч перетворюється на структурований набір подій, який може бути

використаний для статистичного аналізу, побудови моделей прогнозування та оцінювання ефективності окремих ігрових епізодів.

Сучасні системи спортивної аналітики здатні автоматично реєструвати передачі, удари по воротах, перехоплення, відбори м'яча, фоли, зміни володіння м'ячем, переміщення гравців та інші події, що виникають під час матчу [2]. Для кожної дії додатково можуть зберігатися часові характеристики, координати на полі, напрямок розвитку атаки, інформація про виконавця дії, а також взаємне розташування гравців у конкретний момент часу. Такий рівень деталізації дозволяє значно глибше досліджувати ігрові процеси та переходити від загальної статистики до аналізу окремих епізодів і закономірностей поведінки команд.

Одним із найбільш поширених форматів представлення футбольної статистики є event-дані. Вони являють собою послідовність окремих подій матчу, для яких фіксуються тип дії, час її виконання та просторові координати на полі. На основі таких даних можливо відстежувати розвиток атак, аналізувати взаємодію між гравцями, оцінювати небезпечність певних ігрових ситуацій та формувати набір ознак для подальшого використання у моделях машинного навчання. Саме event-дані стали основою для розвитку сучасних систем футбольної аналітики, оскільки дозволяють поєднувати статистичний аналіз із методами інтелектуальної обробки даних.

Приклади використання event-даних для візуалізації та аналізу ігрових дій наведено на рисунку 1.1.

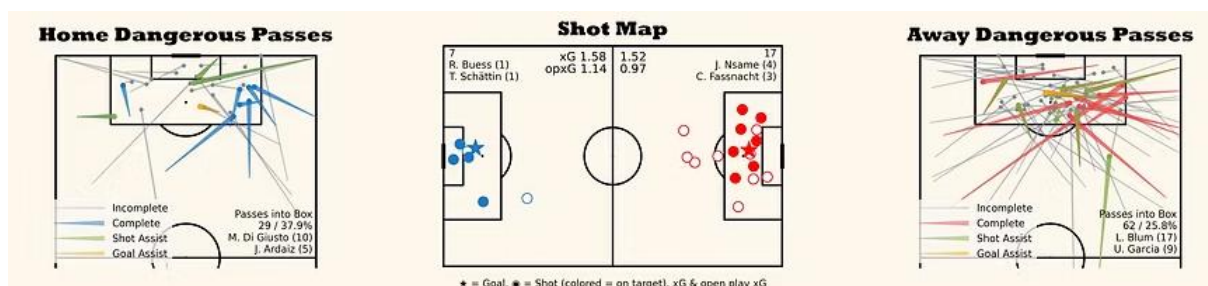


Рисунок 1.1 – Візуалізація футбольних event-даних [3]

У межах футбольної аналітики особливу увагу приділяють результативним діям, оскільки саме вони безпосередньо впливають на створення гольових моментів та підсумковий результат матчу. До таких дій зазвичай відносять удари по воротах, гольові передачі, небезпечні передачі у штрафний майданчик, швидкі атакуювальні комбінації, виходи у перспективні зони атаки та інші епізоди, що потенційно можуть завершитися взяттям воріт. Аналіз подібних дій дозволяє оцінювати не лише загальну активність команди в атаці, а й реальну ефективність створюваних моментів.

Серед усіх атакуювальних дій найбільш важливим об'єктом аналізу є саме удари по воротах, оскільки вони виступають завершальною фазою більшості атак та безпосередньо пов'язані з імовірністю забитого гола [4]. Навіть за високого рівня володіння м'ячем або великої кількості передач команда може не створювати достатньо небезпечних моментів, тоді як якісний удар часто стає вирішальним фактором результативності. Саме тому в сучасній футбольній аналітиці значна частина моделей та статистичних метрик орієнтована на дослідження ударів, оцінювання їх небезпечності та прогнозування ймовірності успішного завершення атакуючого епізоду.

Важливість аналізу ударів пояснюється також тим, що вони характеризуються значною варіативністю ігрових ситуацій. Результат удару залежить не лише від технічних дій гравця, а й від розташування суперників, динаміки розвитку атаки, просторового положення на полі та інших факторів, які складно оцінити за допомогою базової статистики. У зв'язку з цим задача аналізу та прогнозування результативності ударів потребує використання більш складних підходів до обробки даних та побудови аналітичних моделей.

Приклад просторового представлення ударів та їх результатів наведено на рисунку 1.2.

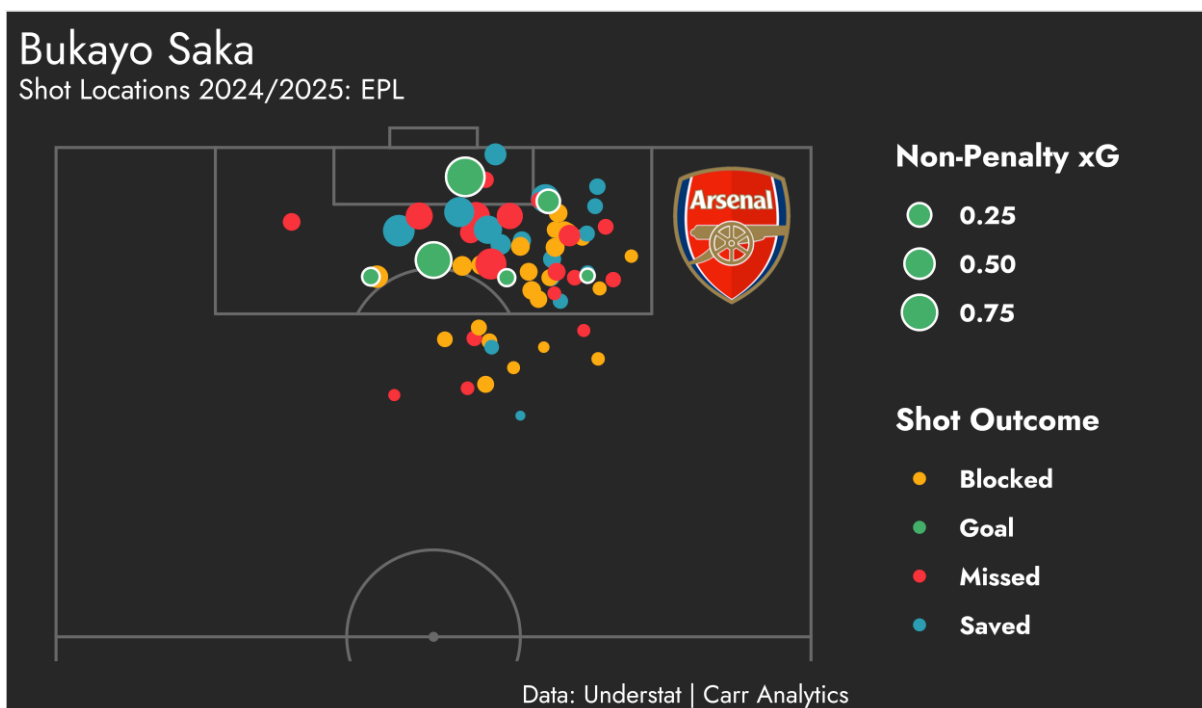


Рисунок 1.2 – Візуалізації ударів та їх результатів у футбольному матчі [5]

Прогнозування результативності ударів у футболі є складною задачею, оскільки успішність завершення атакуючого епізоду залежить від великої кількості взаємопов'язаних факторів. Навіть удари, виконані з подібної позиції на полі, можуть мати зовсім різну ймовірність взяття воріт залежно від розвитку атаки, розташування захисників або дій воротаря. У зв'язку з цим оцінювання небезпечності ударів не може базуватися лише на окремих статистичних показниках, а потребує комплексного аналізу ігрової ситуації.

Одним із ключових факторів є просторове положення гравця у момент виконання удару. На ймовірність забитого гола безпосередньо впливають дистанція до воріт та кут виконання удару, оскільки удари з центральних зон поблизу штрафного майданчика зазвичай мають значно вищу результативність порівняно з ударами з дальньої дистанції або гострих кутів. Важливу роль також відіграє тип атаки, у межах якої було створено момент. Наприклад, швидкі контратаки часто формують більш небезпечні

ситуації через недостатню організацію захисту суперника, тоді як позиційні атаки можуть супроводжуватися високою щільністю оборони [6].

Додатково на результативність удару впливають технічні та ситуаційні характеристики епізоду. До них належать частина тіла, якою виконується удар, швидкість розвитку атаки, кількість захисників у зоні удару, рівень тиску з боку суперника, розташування воротаря та наявність вільного простору перед воротами. У багатьох випадках навіть незначна зміна положення захисника або траєкторії передачі перед ударом може суттєво впливати на підсумковий результат епізоду.

Приклад просторових факторів, що впливають на розвиток атакуючого епізоду, наведено на рисунку 1.3.

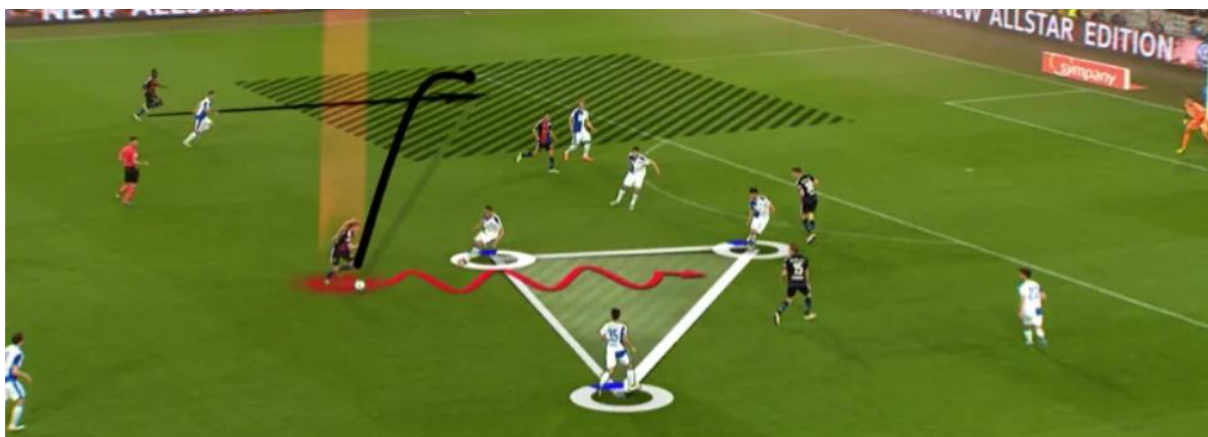


Рисунок 1.3 – Фактори, що впливають на результативність удару [7]

Таким чином, задача прогнозування результативності ударів має виражений багатофакторний характер. Велика кількість параметрів, їх складна взаємодія та залежність від контексту ігрової ситуації ускладнюють побудову універсальних правил оцінювання небезпечності атак. Саме тому сучасна футбольна аналітика дедалі частіше використовує методи машинного навчання, здатні виявляти приховані закономірності у великих обсягах спортивних даних.

Однією з ключових переваг машинного навчання є здатність працювати з багатовимірними даними, які містять велику кількість взаємопов'язаних характеристик. У футбольній аналітиці це дозволяє одночасно враховувати просторові координати гравців, параметри ударів, швидкість розвитку атаки, тип передач, розташування захисників та інші фактори, що впливають на результат ігрового епізоду. Завдяки цьому моделі машинного навчання можуть виявляти приховані залежності між параметрами, які складно або неможливо визначити за допомогою традиційного статистичного аналізу.

Методи машинного навчання широко застосовуються для прогнозування результативності ударів, оцінювання ймовірності забитого гола, аналізу тактичних дій команд та класифікації ігрових ситуацій. Наприклад, моделі можуть використовуватися для автоматичного визначення небезпечних атак, оцінювання ефективності певних тактичних схем або прогнозування успішності дій конкретного гравця у різних ігрових умовах. Крім цього, сучасні алгоритми дозволяють аналізувати великі обсяги історичних даних та формувати моделі, здатні адаптуватися до змін у стилі гри команд або окремих футболістів [8].

Використання машинного навчання у спортивній аналітиці також пов'язане з розвитком методів комп'ютерного бачення та автоматизованого аналізу відео. Це дозволяє не лише працювати зі статистичними таблицями, а й аналізувати безпосередньо відеопотоки матчів, відстежувати переміщення гравців та автоматично формувати набір ознак для подальшого прогнозування. У результаті методи машинного навчання поступово стають одним із основних інструментів сучасної футбольної аналітики, забезпечуючи більш точний та комплексний аналіз результативних дій.

Прогнозування результативних дій має важливе практичне значення для сучасного футболу та активно використовується у різних напрямках спортивної аналітики. Аналіз ймовірності успішного завершення атакуювальних епізодів дозволяє більш об'єктивно оцінювати ефективність

командної гри, визначати сильні та слабкі сторони тактичних схем, а також досліджувати особливості поведінки команди у різних ігрових ситуаціях. На основі подібного аналізу тренерський штаб може отримувати додаткову інформацію щодо якості створюваних моментів, ефективності пресингу, структури атак та рівня реалізації голевих нагод.

Важливим напрямом використання прогнозних моделей є оцінювання індивідуальної ефективності футболістів [9]. Традиційні статистичні показники не завжди дозволяють повністю оцінити внесок гравця у розвиток атак або створення небезпечних моментів. Використання аналітичних моделей дає можливість враховувати контекст ігрових дій, складність ударів, якість передач та інші характеристики епізодів. Це дозволяє більш точно оцінювати атакувальний потенціал гравців, стабільність їхніх дій та ефективність прийняття рішень у різних ситуаціях на полі.

Методи прогнозування результативності також активно застосовуються для підтримки тренерських рішень та підготовки до матчів. Аналіз історичних даних і закономірностей поведінки суперника дозволяє визначати найбільш небезпечні зони атаки, характерні сценарії розвитку ігрових епізодів та типові помилки в обороні. На основі такої інформації можуть коригуватися тактичні установки, структура пресингу або стратегія ведення гри проти конкретного суперника. Крім цього, аналітичні системи використовуються у процесах скаутингу та пошуку нових гравців, оскільки дозволяють оцінювати ефективність футболістів не лише за загальною статистикою, а й за якістю виконуваних атакувальних дій та здатністю створювати небезпечні моменти.

Таким чином, прогнозування результативних дій поступово стає важливою складовою сучасної футбольної аналітики. Використання інтелектуальних методів аналізу даних дозволяє підвищити об'єктивність оцінювання ігрових ситуацій та забезпечує додаткову інформаційну підтримку для прийняття рішень у професійному футболі.

1.2 Аналіз існуючих методів та систем спортивної аналітики у футболі

У роботі [10] досліджено застосування методів машинного навчання для прогнозування результатів футбольних матчів на основі статистичних даних. Для класифікації результатів матчів використовувалися алгоритми Naïve Bayes, Decision Trees, Random Forest та Gradient Boosted Trees. Особливу увагу приділено ансамблевим методам, які краще працюють із неповними даними та забезпечують вищу точність прогнозування. Загальну схему побудови прогнозної моделі наведено на рисунку 1.4.

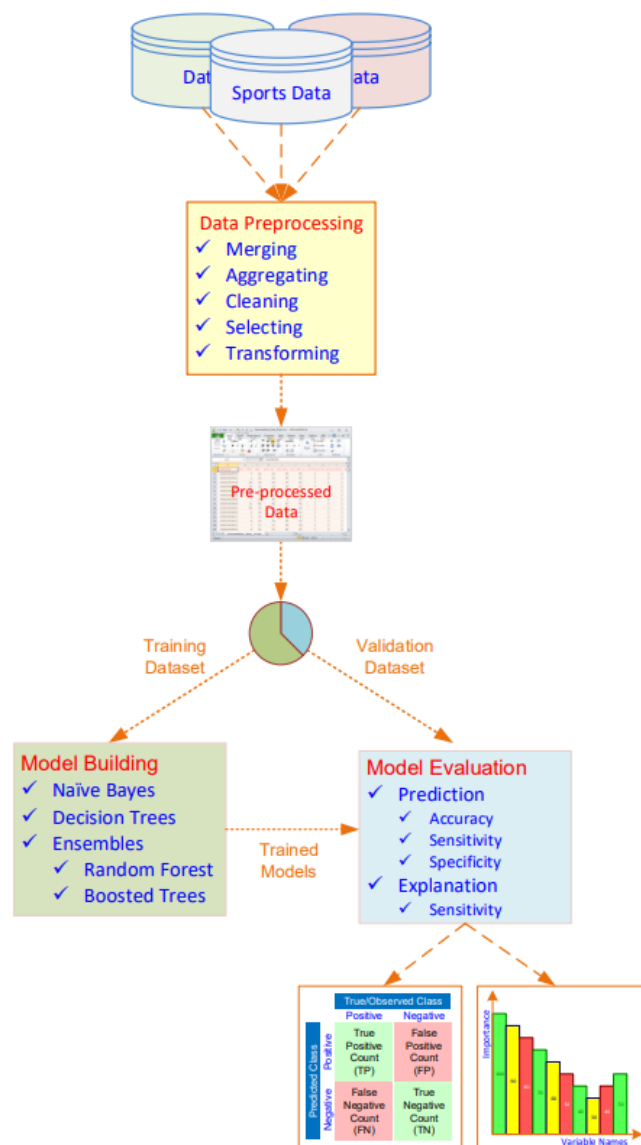


Рисунок 1.4 – Загальна схема процесу побудови прогнозної моделі [10]

За результатами дослідження встановлено, що ансамблеві алгоритми продемонстрували кращу ефективність порівняно з окремими моделями класифікації. Зокрема, модель Random Forest досягла точності 86.3%, а Gradient Boosted Trees – 86.4%.

У роботі [11] досліджено використання методів машинного навчання та ансамблевих алгоритмів для прогнозування результатів футбольних матчів на основі статистичних даних п'яти футбольних клубів. Для побудови моделей застосовувалися алгоритми Naive Bayes, Artificial Neural Networks, K-Nearest Neighbor, C4.5, Random Forest та REPTree. Окрему увагу приділено ансамблевим методам AdaBoost і Bagging, які використовувалися для підвищення точності класифікації.

Для відбору найбільш інформативних ознак застосовано алгоритм Symmetrical Uncertainty. У результаті аналізу встановлено, що найбільший вплив на результат матчу мають кількість пропущених голів, результат першого тайму, команда, яка першою відкрила рахунок, та точність ударів. Це підтверджує важливість результативних дій у задачах футбольної аналітики.

Отримані результати показали перевагу ансамблевих методів над окремими алгоритмами класифікації. Найкращу ефективність продемонстрували моделі AdaBoost та Artificial Neural Networks. Для футбольного клубу Barcelona точність класифікації досягла 99.3%.

У роботі [9] досліджено використання методів машинного навчання для автоматичного оцінювання ефективності футбольних гравців на основі статистичних характеристик ігрових дій. Для розв'язання задачі запропоновано модель PEMLM (Performance Evaluation Machine Learning Model), схема якої наведена на рисунку 1.5.

У дослідженні використовувалися два набори даних, що містили інформацію про матчеві та тренувальні сесії. Після попередньої обробки та кластеризації даних виконувалася класифікація гравців за трьома категоріями: active, normal та weak. Для побудови моделей застосовувалися

алгоритми Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, Gaussian Naïve Bayes, Multi-Layer Perceptron та K-Nearest Neighbor.

За результатами експериментів найкращу ефективність для матчевих даних продемонстрував алгоритм Decision Tree із точністю 100%, тоді як для тренувальних даних найкращі результати отримано методом K-Nearest Neighbor із точністю 96%. Для перевірки стійкості моделей та оцінювання перенавчання використовувалася k-fold cross-validation.

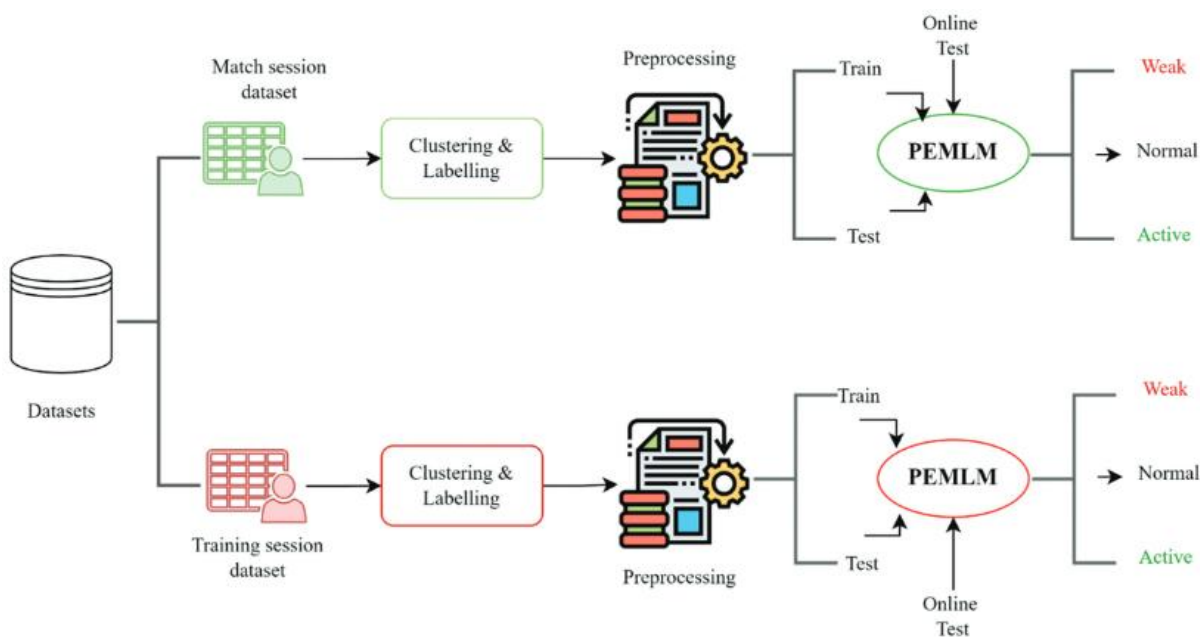


Рисунок 1.5 – Схема моделі PEMLM для оцінювання ефективності футболістих гравців [9]

У роботі [8] досліджено використання методів машинного навчання для оцінювання ефективності команд та підтримки стратегічних рішень у футболі. Основну увагу приділено аналізу результативності атакуючих дій, прогнозуванню небезпечності моментів та оцінюванню впливу просторових взаємодій гравців на розвиток атаки.

У дослідженні використовувалися набори даних Wyscout та Metrica. Event-дані Wyscout містили інформацію про матчі провідних європейських ліг, тоді як tracking-дані Metrica забезпечували просторову інформацію про

розташування м'яча та гравців із частотою 25 кадрів на секунду. Для аналізу атакуювальних дій застосовувалися моделі expected goals (xG), expected threat (xT) та expected possession value (EPV), які дозволяють оцінювати не лише ймовірність забитого гола, а й потенційну небезпечність розвитку атаки. Приклад використання моделі expected possession value наведено на рисунку 1.6.

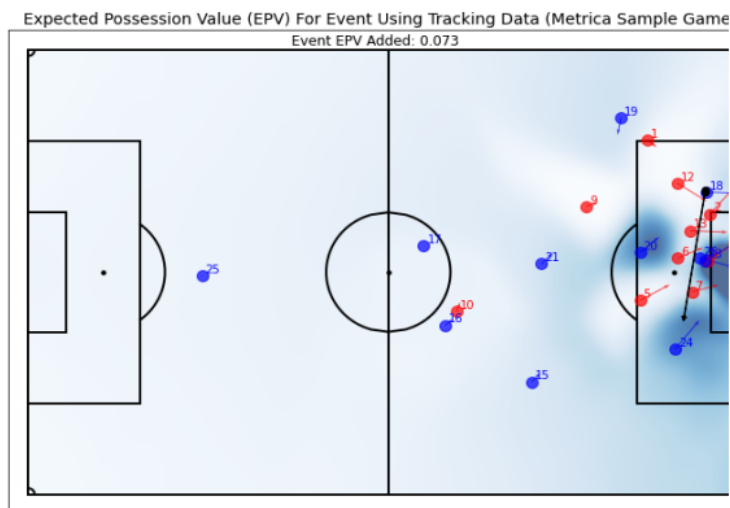


Рисунок 1.6 – Приклад використання моделі expected possession value для аналізу варіантів розвитку атаки [8]

Для побудови моделей використовувалися алгоритми Logistic Regression, Random Forest та Extreme Gradient Boosting. Найкращі результати продемонстрував алгоритм Extreme Gradient Boosting, який досяг значення AUC 0.8 та середньоквадратичної помилки 0.28. Отримані результати підтвердили ефективність ансамблевих ML-моделей для задач футбольної аналітики та прогнозування результативних дій.

Особливістю дослідження є комплексне поєднання event-даних, tracking-даних та просторових моделей оцінювання атакуювальних дій, що дозволяє аналізувати не лише завершальні удари, а й процес створення небезпечних моментів.

У роботі [12] досліджено підвищення якості прогнозування результативності ударів у футболі за допомогою feature engineering та методів машинного навчання. Основною задачею було побудова моделі expected goals (xG), здатної оцінювати ймовірність взяття воріт на основі характеристик атакуючого епізоду.

У дослідженні зазначено, що традиційні статистичні показники, зокрема кількість ударів або володіння м'ячем, не забезпечують повного оцінювання якості створених моментів. У зв'язку з цим застосовано підхід expected goals, у межах якого кожному удару призначається ймовірність забитого гола залежно від параметрів ігрової ситуації.

Для побудови xG-моделі аналізувалися просторові та контекстні характеристики ударів, серед яких координати гравця, дистанція до воріт, кут удару та тип атакуючого епізоду. Особливу увагу приділено feature engineering, що дозволяє формувати більш інформативні ознаки для моделей машинного навчання.

Отримані результати показали, що використання додаткових просторових і контекстних характеристик суттєво покращує якість прогнозування результативності ударів порівняно з базовими статистичними моделями. Особливістю роботи є орієнтація не лише на підвищення точності прогнозування, а й на забезпечення інтерпретованості результатів за рахунок використання зрозумілих та обґрунтованих ознак.

У роботі [6] запропоновано модель expected goals для оцінювання ймовірності взяття воріт на основі синхронізованих positional- та event-даних. Основною метою дослідження було підвищення точності оцінювання якості ударів та ефективності атакуючих дій порівняно з традиційними статистичними підходами.

У дослідженні зазначено, що через невелику кількість голів у футболі саме удари часто використовуються як основний показник атакуючої активності. Водночас результативність удару суттєво залежить від просторового контексту ігрового епізоду. Для врахування цих факторів

використовувалися синхронізовані positional- та event-дані, що дозволило поєднати інформацію про координати гравців і характеристики ігрових подій.

У межах роботи було проаналізовано 105 627 ударів із матчів німецької Бундесліги. Для побудови xG-моделі застосовано алгоритм Extreme Gradient Boosting із набором hand-crafted ознак, сформованих на основі просторових та контекстних характеристик ударів. Серед ознак враховувалися дистанція до воріт, кут удару, розташування захисників, положення воротаря та динаміка розвитку атаки. Приклад синхронізації positional- і event-даних наведено на рисунку 1.7.

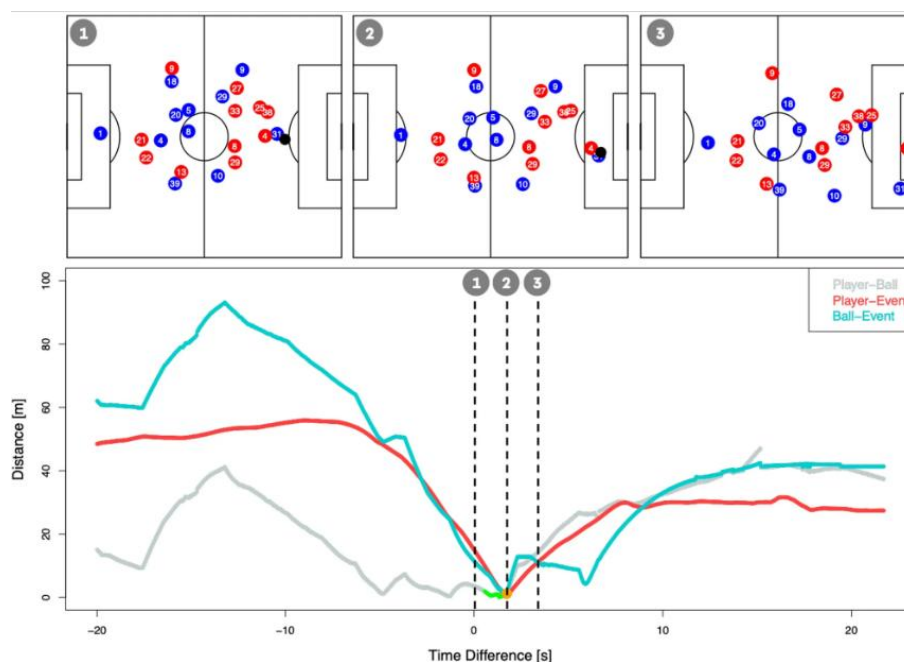


Рисунок 1.7 – Приклад синхронізації positional- та event-даних для аналізу ударів у футболі [6]

Отримані результати показали високу ефективність запропонованої моделі. Встановлено, що команда з вищим значенням xG перемогала у 73.3% матчів без нічиїх. Крім цього, модель дозволила оцінювати якість голевих передач незалежно від завершальної дії нападника.

Особливістю дослідження є поєднання positional- та event-даних, що забезпечує більш деталізований аналіз ігрових епізодів та дозволяє враховувати просторові взаємодії між гравцями під час розвитку атаки.

У роботі [13] запропоновано підхід до автоматичного виявлення футбольних подій на основі tracking-даних. На відміну від підходів, що використовують лише статистичні показники або окремі зображення ігрових епізодів, у дослідженні аналіз базувався на просторових координатах гравців і м'яча, отриманих у процесі відстеження матчу.

Основною задачею було автоматичне визначення ігрових подій, зокрема передач, ударів, перехоплень, прийомів м'яча та стандартних положень без необхідності ручної розмітки матчів. Для цього використовувався двоетапний алгоритм, структура якого наведена на рисунку 1.8.

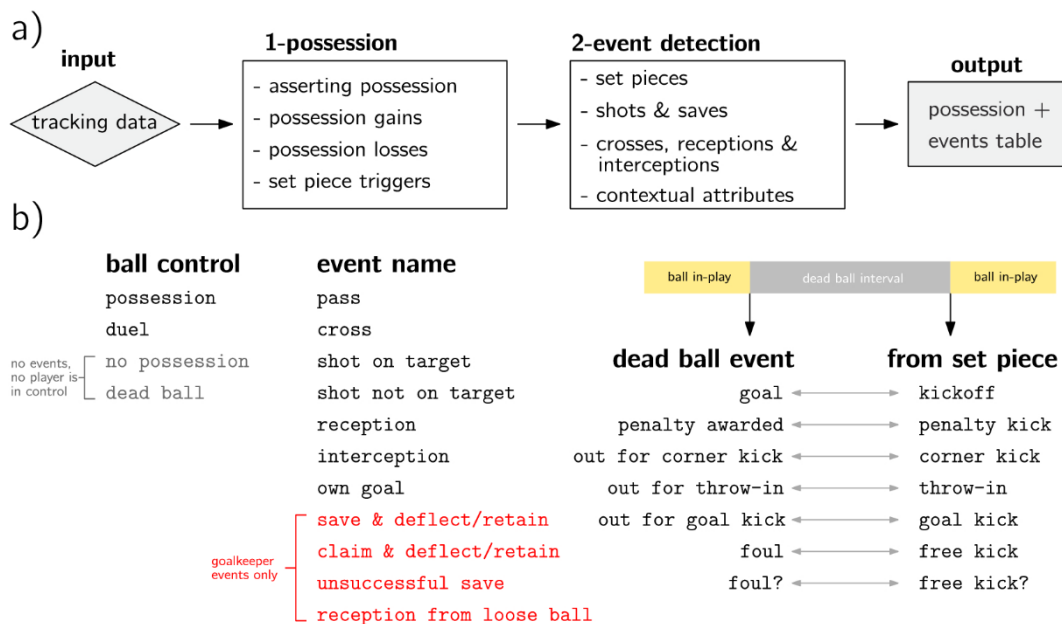


Рисунок 1.8 – Обчислювальна схема автоматичного виявлення футбольних подій [13]

На першому етапі визначалося володіння м'ячем та аналізувалися конфігурації гравців у моменти, коли м'яч не перебував у грі. Отримана

інформація використовувалася для виявлення стандартних положень і формування контексту ігрових ситуацій. На другому етапі застосовувався детермінований алгоритм на основі дерева рішень, який використовував зміни володіння м'ячем та правила футболу для автоматичного визначення типу події. У результаті формувалася таблиця подій із часовими, просторовими та контекстними характеристиками.

Особливістю підходу є використання tracking-даних як основного джерела інформації, що дозволяє враховувати просторовий контекст ігрових епізодів, взаємне розташування футболістів та динаміку руху м'яча. Отримані результати продемонстрували високу ефективність автоматичного виявлення футбольних подій для різних турнірів і постачальників tracking-даних. Також показано, що використання додаткових джерел інформації, зокрема відеоданих або відстеження положення частин тіла гравців, дозволяє автоматично визначати складніші ігрові події, наприклад дриблінг або боротьбу за м'яч.

У роботі [14] запропоновано підхід до автоматичної класифікації футбольних подій на основі методів глибокого навчання та комп'ютерного бачення. На відміну від підходів, що використовують статистичні дані матчів, у дослідженні аналіз виконувався на основі зображень ігрових епізодів, отриманих із футбольних трансляцій.

Основною задачею було автоматичне визначення типу футбольної події, зокрема передачі, удару, штрафного удару, підкату або кутового. Для цього використовувалася модель, що поєднує згортковий автоенкодер (Convolutional Autoencoder) та багатошаровий класифікатор Extreme Learning Machine. Архітектуру запропонованої моделі наведено на рисунку 1.9.

У межах запропонованого підходу згортковий автоенкодер використовувався для автоматичного виділення ознак і зменшення розмірності вхідних даних. Після етапу кодування сформовані ознаки передавалися до багатошарової моделі Extreme Learning Machine, яка

виконувала класифікацію футбольних подій. Використання автоенкодера дозволило зменшити обсяг обчислень та підвищити швидкість обробки даних, що є важливим для аналізу спортивних відео у режимі реального часу.

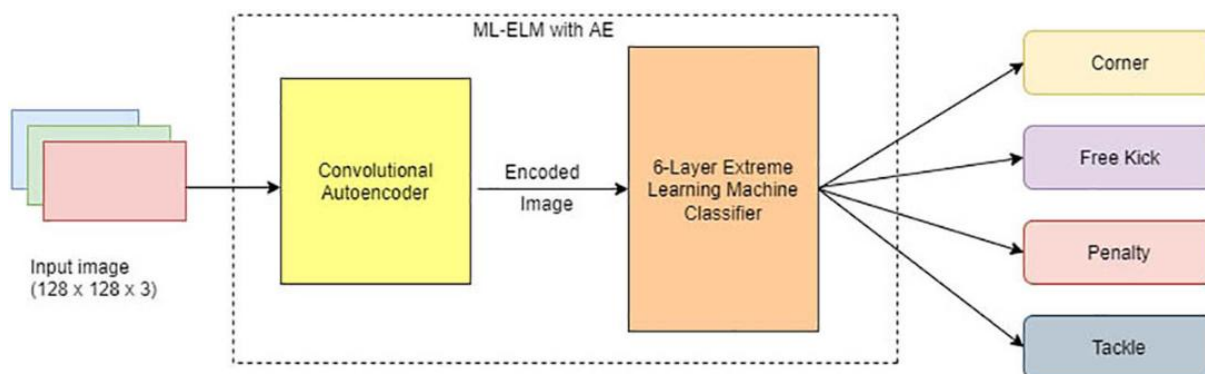


Рисунок 1.9 – Архітектура моделі класифікації футбольних подій [14]

Отримані результати показали, що поєднання згорткового автоенкодера та багатошарового Extreme Learning Machine забезпечує вищу точність класифікації порівняно з традиційними підходами. Також показано, що методи глибокого навчання дозволяють автоматично формувати інформативні ознаки без необхідності ручного визначення характеристик ігрових епізодів.

У роботі [15] досліджено використання методів глибокого навчання для аналізу тактичних дій у футболі. Основну увагу приділено аналізу просторово-часових закономірностей поведінки гравців, командних взаємодій та прогнозуванню розвитку ігрових ситуацій на основі tracking- і event-даних.

У дослідженні зазначено, що традиційні статистичні методи не дозволяють повністю описати складні багатокористувацькі взаємодії, характерні для футбольних матчів. Для вирішення цієї проблеми запропоновано використання глибоких нейронних архітектур, здатних автоматично аналізувати просторові та часові залежності між діями гравців.

Основну увагу приділено моделі Tactiformer – ієрархічній transformer-архітектурі, адаптованій для задач футбольної аналітики. Архітектуру моделі наведено на рисунку 1.10.

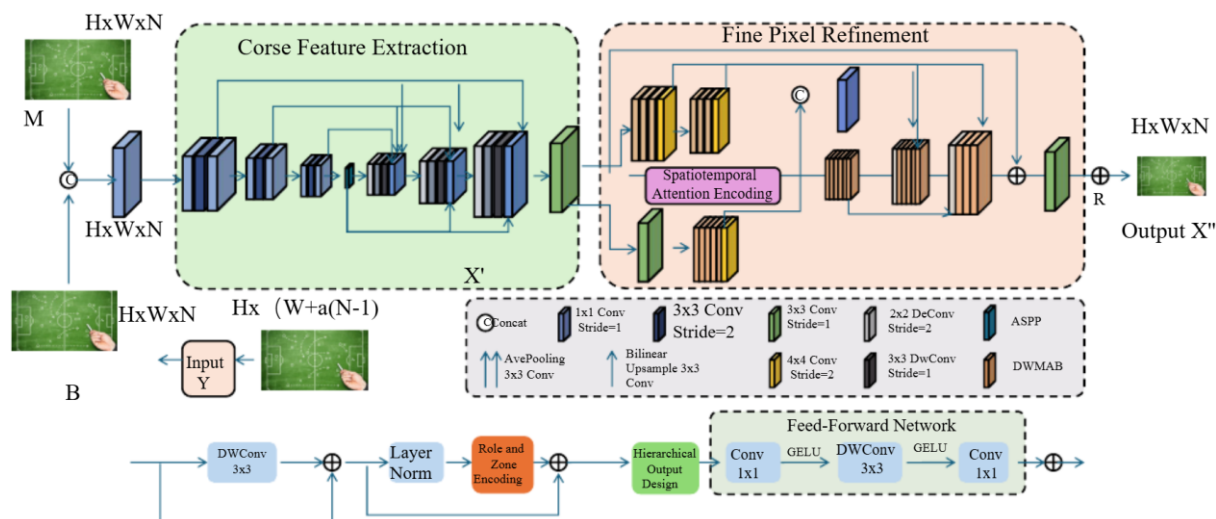


Рисунок 1.10 – Архітектура моделі Tactiformer для аналізу футбольних матчів [15]

Модель складається з декількох рівнів аналізу та використовує механізми spatiotemporal attention, role-aware embedding і contextual aggregation для врахування просторового розташування гравців, їх ролей та динаміки розвитку ігрових ситуацій. Крім цього, архітектура підтримує аналіз тактичних шаблонів, прогнозування розвитку атак та побудову теплових карт активності команд.

У роботі також розглядається модель StratGaze, яка використовує механізми стратегічної уваги для автоматичного виявлення тактичних мотивів і прогнозування траєкторій розвитку гри. Отримані результати підтвердили перспективність використання transformer-архітектур для задач тактичного аналізу та прогнозування атаквальних дій у футболі. Використання великих наборів tracking-даних дозволило підвищити деталізацію аналізу та точність моделювання ігрових ситуацій.

У роботі [16] запропоновано інтелектуальну систему для автоматизованої анотації футбольних матчів із використанням методів штучного інтелекту та сенсорних даних. Основною метою дослідження є зменшення складності та вартості ручної розмітки футбольних подій, необхідної для побудови сучасних систем спортивної аналітики та навчання ML-моделей.

У дослідженні зазначено, що більшість існуючих систем анотації футбольних матчів базуються переважно на відеоаналізі та методах комп'ютерного бачення, що ускладнює аналіз командної динаміки та повне відстеження всіх гравців на полі. Для вирішення цієї проблеми у системі FootApp використовується поєднання відеоданих, голосової та сенсорної інформації. Загальну структуру системи наведено на рисунку 1.11.

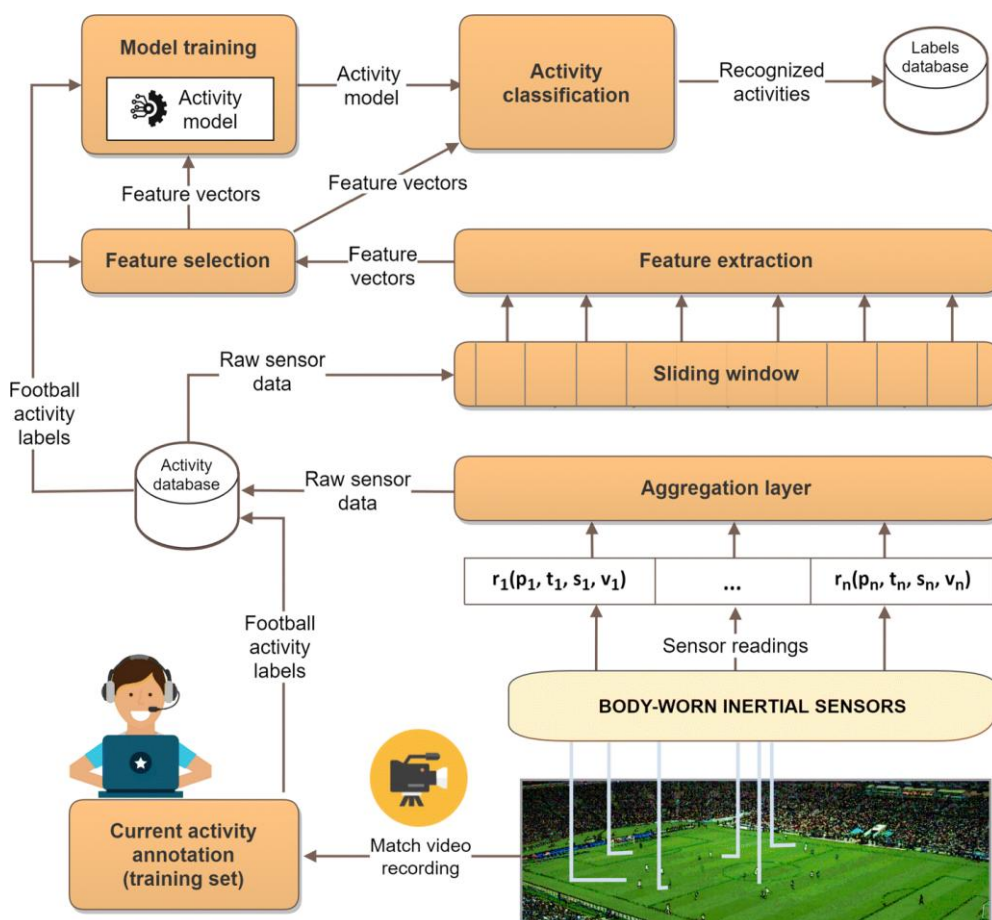


Рисунок 1.11 – Архітектура системи FootApp для автоматизованої анотації футбольних матчів [16]

У межах системи використовуються інерціальні сенсори, закріплені на гравцях, які формують потік даних про рухову активність футболістів. Після агрегації даних, формування часових вікон та виділення ознак застосовуються алгоритми машинного навчання для класифікації футбольних дій. Додатково система підтримує інтерактивну анотацію матчів за допомогою голосового та сенсорного інтерфейсу.

Особливістю підходу є використання методів штучного інтелекту для автоматичної перевірки узгодженості анотацій та виявлення помилок у розмітці. Це дозволяє підвищити якість формування навчальних наборів даних для задач прогнозування результативних дій та спортивної аналітики. Проведене експериментальне оцінювання підтвердило ефективність запропонованої системи для автоматизованого збору та аналізу футбольних даних.

У роботі [17] запропоновано відкриту платформу для аналізу просторово-часових футбольних даних, орієнтовану на прогнозування ігрових подій та підтримку задач спортивної аналітики. Основною метою дослідження є створення універсального середовища для обробки event- і tracking-даних із використанням методів машинного навчання та reinforcement learning.

У дослідженні зазначено, що обмежена доступність event-даних та висока вартість tracking-даних ускладнюють побудову комплексних систем аналізу матчів. Для вирішення цієї проблеми запропоновано платформу OpenSTARLab, архітектуру якої наведено на рисунку 1.12.

Платформа складається з декількох компонентів. Pre-processing Package забезпечує стандартизацію та інтеграцію event- і tracking-даних у єдині формати представлення. Event Modeling Package реалізує моделі прогнозування футбольних подій на основі методів глибокого навчання, а RLearn Package використовується для задач reinforcement learning та моделювання ігрових сценаріїв. Крім цього, система підтримує формування

просторово-часових представлень дій гравців, аналіз траєкторій руху та симуляцію розвитку матчів.

У межах дослідження проводилися експерименти з прогнозування типів ігрових дій та часу виникнення подій. Отримані результати показали високу точність моделей прогнозування та стабільність симуляції розвитку ігрових епізодів. Також продемонстровано можливість використання reinforcement learning для аналізу стратегічних рішень і моделювання поведінки команд у різних ситуаціях на полі.

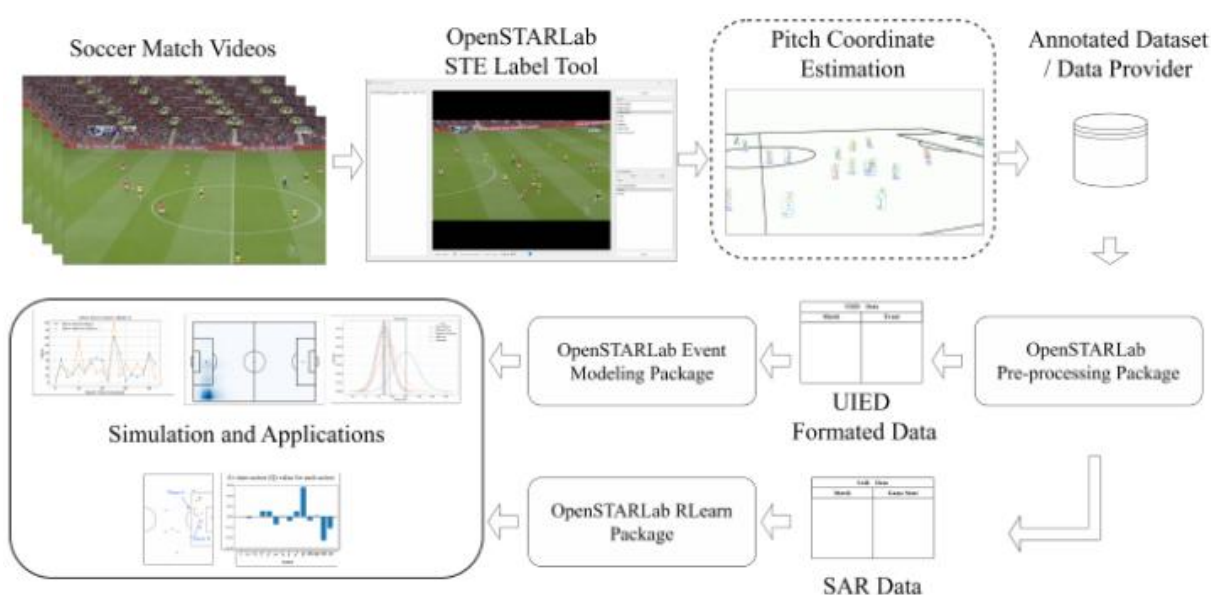


Рисунок 1.12 – Архітектура платформи OpenSTARLab для аналізу футбольних даних [17]

Особливістю OpenSTARLab є орієнтація на створення комплексного середовища для спортивної аналітики, що підтримує інтеграцію різних типів футбольних даних, візуалізацію результатів аналізу та виконання задач тактичного аналізу і прогнозування результативних дій.

Проведений аналіз існуючих методів і систем прогнозування результативних дій у футболі показав, що сучасна спортивна аналітика активно переходить від використання базових статистичних показників до комплексного аналізу просторових, часових та контекстних характеристик

ігрових епізодів. Якщо ранні підходи переважно базувалися на класичних алгоритмах машинного навчання та загальній статистиці матчів, то сучасні дослідження дедалі частіше використовують event- і tracking-дані, моделі expected goals, transformer-архітектури та методи глибокого навчання для аналізу взаємодії гравців і прогнозування розвитку атакуючих ситуацій.

Особливу увагу у сучасних дослідженнях приділено оцінюванню якості ударів та атакуючих дій за допомогою моделей expected goals і просторового аналізу. Використання positional-даних дозволяє враховувати розташування гравців, динаміку руху м'яча та контекст розвитку атаки, що суттєво підвищує якість прогнозування результативності порівняно з традиційними статистичними підходами. Використання методів глибокого навчання та transformer-моделей також забезпечує можливість автоматичного виявлення складних закономірностей у поведінці команд та аналізу багатокористувацьких взаємодій на полі.

Попри це, існуючі підходи мають низку обмежень. Частина моделей потребує великих обсягів високоякісних tracking-даних, отримання яких є складним і дорогим процесом. Інші підходи характеризуються високою обчислювальною складністю або недостатньою інтерпретованістю результатів. Навіть сучасні моделі не завжди повною мірою враховують тактичний контекст ігрових ситуацій та взаємозв'язки між діями всіх учасників епізоду.

Таким чином, сучасні підходи до футбольної аналітики демонструють ефективність використання методів машинного навчання, просторового аналізу та deep learning для дослідження результативних дій у футболі. Водночас складність ігрових взаємодій, залежність результативності від контексту епізоду та необхідність обробки великих обсягів event- і tracking-даних свідчать про актуальність подальших досліджень у напрямі інтелектуального аналізу спортивних даних.

1.3 Постановка задачі дослідження

Зростання обсягів спортивних даних та розвиток сучасних засобів спортивної аналітики створюють передумови для застосування інтелектуальних методів аналізу і прогнозування у футболі. Використання event-даних футбольних матчів дозволяє детально описувати ігрові ситуації, враховувати характеристики атакуювальних дій та формувати моделі, здатні оцінювати ймовірність результативного завершення певних епізодів. При цьому ефективність прогнозування значною мірою залежить від якості підготовки даних, вибору інформативних ознак та використаних методів машинного навчання.

Сучасні системи спортивної аналітики активно використовують статистичні методи, моделі машинного навчання та інтелектуальний аналіз даних для оцінювання результативності ігрових дій. При цьому задачі прогнозування результативних дій залишаються складними через багатофакторність футбольних подій, залежність результату від контексту ігрової ситуації та необхідність обробки значних обсягів неоднорідних даних. Додаткову складність створює потреба врахування просторових і часових характеристик атакуювальних епізодів, що безпосередньо впливають на результативність ударів. У зв'язку з цим актуальною є задача розробки інтелектуальної системи, здатної автоматизувати аналіз спортивних даних та прогнозування результативних дій на основі event-даних футбольних матчів.

Метою роботи є розробка інтелектуальної системи прогнозування результативності ударів у футбольній аналітиці на основі методів машинного навчання.

Для досягнення поставленої мети необхідно виконати такі задачі:

- проаналізувати особливості спортивної аналітики та сучасні підходи до прогнозування результативних дій у футболі;

- дослідити методи машинного навчання, що застосовуються у задачах аналізу спортивних даних;
- виконати формування та попередню обробку набору даних на основі подій футбольних матчів;
- реалізувати моделі машинного навчання для прогнозування результативності ударів;
- провести експериментальне дослідження та порівняльний аналіз якості моделей прогнозування;
- розробити програмну реалізацію інтелектуальної системи прогнозування результативності ударів;
- перевірити роботу розробленої системи на реальних і змодельованих ігрових ситуаціях.

2 МЕТОДИ ТА МОДЕЛІ ПРОГНОЗУВАННЯ РЕЗУЛЬТАТИВНИХ ДІЙ У СПОРТИВНІЙ АНАЛІТИЦІ

Сучасні системи спортивної аналітики базуються на використанні великої кількості статистичних, просторових і часових даних, отриманих під час футбольних матчів. Для прогнозування результативних дій застосовуються як класичні алгоритми машинного навчання, так і сучасні архітектури глибокого навчання, здатні враховувати складні залежності між положенням гравців, розвитком атаки та контекстом ігрового епізоду. Важливу роль при цьому відіграють методи обробки event- і tracking-даних, способи формування ознак, а також підходи до оцінювання якості побудованих моделей.

2.1 Класичні статистичні методи аналізу спортивних даних

Статистичний аналіз є одним із базових підходів до дослідження даних та використовується для виявлення закономірностей, оцінювання характеристик вибірки й аналізу залежностей між окремими показниками. Основною задачею статистичних методів є перетворення великого обсягу числових даних у набір узагальнених характеристик, що дозволяють оцінити структуру даних, розподіл значень та особливості поведінки досліджуваних показників.

У задачах спортивної аналітики статистичні методи застосовуються для аналізу кількісних характеристик матчів, ігрових дій та індивідуальних показників гравців. Особливістю спортивних даних є їх багатофакторність, нерівномірний розподіл та наявність значної кількості взаємопов'язаних параметрів. У зв'язку з цим статистичний аналіз використовується як для первинного дослідження даних і формування ознак, так і для оцінювання залежностей між характеристиками ігрових епізодів та результативністю команд або окремих дій.

Методи описової статистики (рисунок 2.1) дозволяють оцінювати основні характеристики вибірки та виконувати узагальнення числових даних. До таких методів належать обчислення середнього значення, медіани, моди, дисперсії та стандартного відхилення. Використання цих показників забезпечує можливість аналізу центральних тенденцій, ступеня варіативності даних та особливостей розподілу окремих характеристик [18].

Середнє значення використовується для визначення узагальненого рівня певного показника у вибірці, наприклад середньої кількості ударів або передач за матч. Медіана дозволяє оцінювати центральне значення розподілу у випадках нерівномірного розподілу даних або наявності аномальних значень. Дисперсія та стандартне відхилення характеризують ступінь відхилення значень відносно середнього та використовуються для оцінювання стабільності статистичних показників.

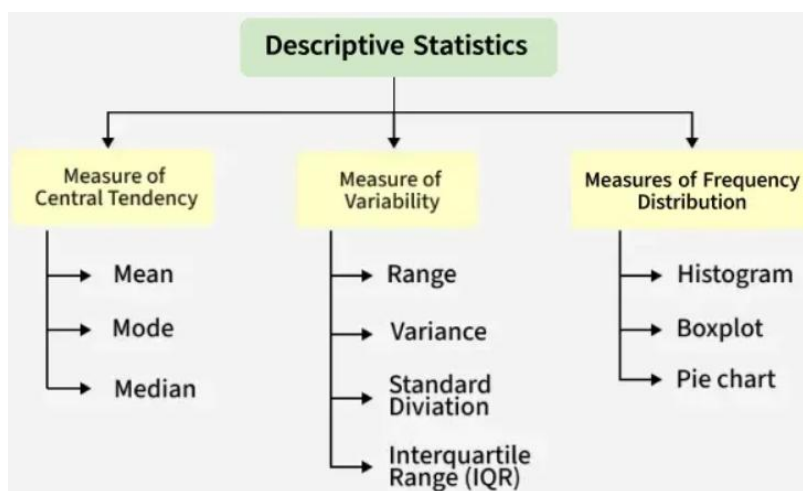


Рисунок 2.1 – Класифікація методів описової статистики [18]

Важливу роль у спортивній аналітиці відіграють відносні статистичні показники, що визначаються як співвідношення окремих характеристик. До таких показників належать точність передач, відсоток виграних єдиноборств, ефективність реалізації ударів та інші метрики, які дозволяють виконувати порівняння незалежно від кількості ігрових дій або проведених

матчів. Використання відносних показників забезпечує більш коректне оцінювання ефективності команд і окремих гравців [19].

Для аналізу структури та розподілу даних застосовуються методи статистичної візуалізації. Гістограми використовуються для оцінювання частоти появи певних значень у вибірці, boxplot-діаграми – для аналізу розподілу даних та виявлення аномальних значень, а теплові карти дозволяють досліджувати просторовий розподіл ігрових дій на полі [3]. Використання таких методів забезпечує більш наочний аналіз статистичних закономірностей та дозволяє виконувати попереднє дослідження даних перед побудовою прогнозних моделей.

Методи кореляційного та регресійного аналізу використовуються для дослідження залежностей між окремими статистичними показниками та оцінювання впливу різних факторів на результат ігрових епізодів. Основною задачею таких методів є виявлення взаємозв'язків між характеристиками даних та побудова математичних моделей, що дозволяють прогнозувати значення певних показників або ймовірність виникнення подій.

Для оцінювання сили зв'язку між двома змінними застосовується кореляційний аналіз. Одним із найбільш поширених методів є коефіцієнт кореляції Пірсона, який дозволяє визначити ступінь лінійної залежності між показниками.

Значення коефіцієнта кореляції знаходиться в діапазоні від -1 до 1 , де значення, близькі до 1 , свідчать про сильний прямий зв'язок, значення, близькі до -1 , – про обернений зв'язок, а значення, близькі до 0 , – про відсутність вираженої залежності. У спортивній аналітиці кореляційний аналіз використовується для дослідження залежностей між кількістю ударів, володінням м'ячем, точністю передач, інтенсивністю атак та результативністю команди.

Для моделювання залежностей між змінними застосовуються методи регресійного аналізу (рисунок 2.2). Лінійна регресія використовується для

прогнозування числових значень на основі одного або декількох факторів та дозволяє оцінити вплив окремих параметрів на результуючий показник. У задачах спортивної аналітики лінійна регресія може використовуватися для прогнозування кількості ударів, очікуваної кількості голів або інших статистичних характеристик матчу [20].

У випадках, коли необхідно прогнозувати ймовірність настання певної події, застосовується логістична регресія. На відміну від лінійної регресії, результат логістичної моделі подається у вигляді ймовірності належності об'єкта до певного класу [21]. У футбольній аналітиці логістична регресія широко використовується для оцінювання ймовірності забиття гола, успішності передачі або реалізації небезпечного моменту залежно від набору статистичних та просторових характеристик ігрового епізоду.

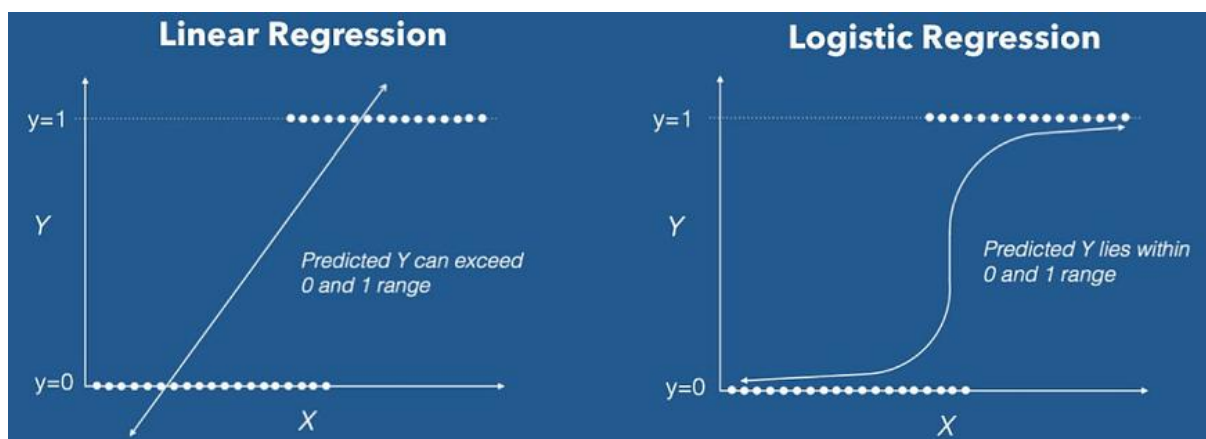


Рисунок 2.2 – Принцип роботи лінійної та логістичної регресії [21]

Використання кореляційного та регресійного аналізу дозволяє виявляти статистично значущі фактори, що впливають на результативність команд і окремих дій. Проте такі методи мають обмеження при аналізі складних нелінійних залежностей та великої кількості взаємопов'язаних параметрів, що зумовлює необхідність використання більш складних алгоритмів машинного навчання.

Імовірнісні статистичні моделі використовуються для оцінювання ймовірності виникнення певних подій на основі статистичних закономірностей у даних. У спортивній аналітиці такі моделі застосовуються для прогнозування результатів матчів, оцінювання ефективності атакуювальних дій та аналізу результативності команд. На відміну від описових статистичних методів, імовірнісні моделі дозволяють не лише аналізувати наявні дані, а й оцінювати ймовірність майбутніх подій на основі виявлених закономірностей.

Однією з найбільш поширених статистичних моделей у футбольній аналітиці є розподіл Пуассона, що використовується для моделювання кількості подій протягом певного проміжку часу. У футболі така модель застосовується для оцінювання ймовірності забиття певної кількості голів за матч, де параметр λ характеризує середню результативність команди [22]. Якщо значення λ є невеликим, найбільш імовірними вважаються рахунки з малою кількістю голів, тоді як зі збільшенням λ розподіл стає більш рівномірним, а ймовірність результативних матчів зростає (рисунок 2.3). Використання розподілу Пуассона дозволяє оцінювати ймовірності різних рахунків матчу та формувати прогноз результативності команд на основі статистичних показників атаки й оборони.

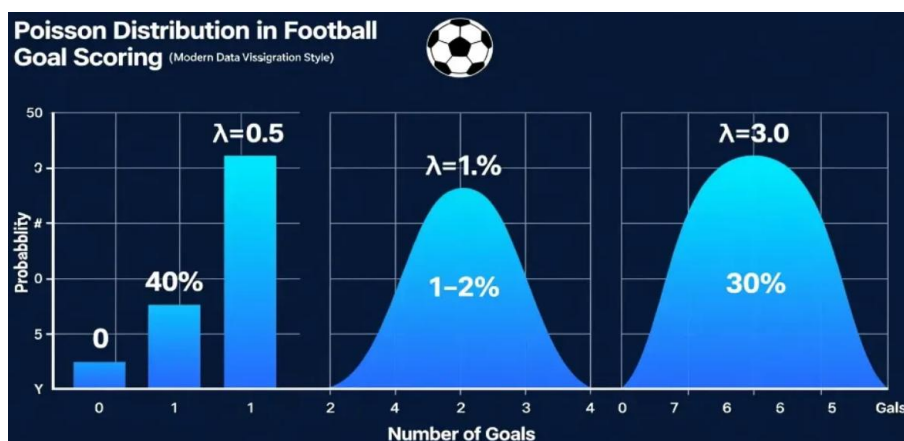


Рисунок 2.3 – Використання розподілу Пуассона для моделювання результативності у футболі

Для побудови моделей прогнозування рахунку використовуються статистичні характеристики команд, зокрема середня кількість забитих і пропущених голів, ефективність атаквальних дій та показники гри в атаці й обороні. На основі таких параметрів визначаються очікувані значення результативності команд, після чого за допомогою імовірнісних розподілів оцінюється ймовірність різних результатів матчу. Подібні підходи активно застосовуються у спортивній аналітиці для прогнозування результатів футбольних зустрічей та аналізу ефективності команд.

Окреме значення у сучасній футбольній аналітиці мають моделі expected goals (xG), що використовуються для оцінювання якості ударів та ймовірності забиття гола у конкретному ігровому епізоді. На відміну від традиційного підрахунку кількості ударів, xG-моделі враховують контекст виконання удару, зокрема відстань до воріт, кут удару, тип передачі, частину тіла, якою виконується удар, положення гравців та інші характеристики епізоду. Результатом роботи моделі є значення ймовірності забиття гола для кожного удару [23].

Використання моделей expected goals дозволяє більш об'єктивно оцінювати ефективність атаквальних дій та якість створених моментів незалежно від фактичного результату епізоду. Такі моделі широко застосовуються для аналізу командної гри, оцінювання ефективності футболістів та побудови систем прогнозування результативних дій. Водночас точність імовірнісних моделей значною мірою залежить від якості даних, кількості врахованих параметрів та здатності моделі описувати складні просторово-часові взаємозв'язки між ігровими подіями.

2.2 Методи машинного навчання у задачах спортивної аналітики

Зростання обсягів футбольних даних та ускладнення сучасної спортивної аналітики сприяли активному впровадженню методів машинного навчання. На відміну від традиційних статистичних підходів, які

здебільшого базуються на заздалегідь визначених показниках та ручному аналізі даних, методи машинного навчання дозволяють автоматично виявляти закономірності у великих масивах інформації та формувати прогностичні моделі на основі реальних ігрових даних. Це створює можливість більш глибокого аналізу футбольних матчів та підвищує об'єктивність оцінювання ігрових ситуацій.

Одним із поширених підходів є використання дерев рішень, що виконують послідовний поділ даних за визначеними умовами. Кожен вузол дерева відповідає перевірці певної ознаки, а кінцеві вузли формують результат класифікації або прогнозування. Перевагою такого підходу є наочність структури моделі та можливість аналізу впливу окремих параметрів на кінцевий результат. Водночас окремі дерева рішень мають схильність до перенавчання та можуть демонструвати нестабільність при зміні навчальної вибірки.

Ще одним поширеним методом є Support Vector Machine (SVM), принцип роботи якого полягає у побудові оптимальної межі розділення між класами у просторі ознак (рисунок 2.4).

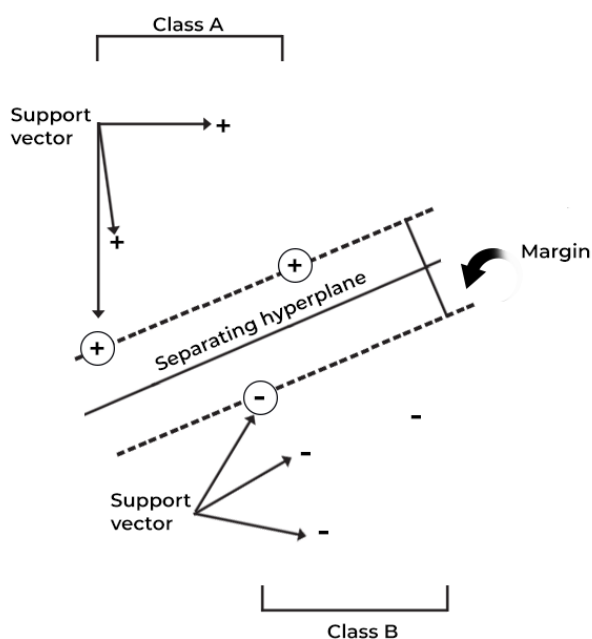


Рисунок 2.4 – Support vector machine [24]

Використання ядерних функцій дозволяє моделі працювати зі складними нелінійними залежностями та виконувати класифікацію багатовимірних даних [24].

У спортивній аналітиці метод SVM може застосовуватися для класифікації ігрових епізодів, аналізу тактичних патернів та прогнозування результативних дій. Недоліком такого підходу є висока обчислювальна складність при роботі з великими наборами даних та необхідність підбору параметрів моделі.

Ансамблеві методи машинного навчання базуються на поєднанні декількох моделей з метою підвищення точності прогнозування та покращення стійкості результатів. Основна ідея таких підходів полягає у зменшенні впливу помилок окремих моделей шляхом комбінування їх прогнозів. У задачах спортивної аналітики ансамблеві алгоритми демонструють високу ефективність при роботі з великою кількістю статистичних, просторових та контекстних ознак.

Одним із поширених підходів є *bagging*, що передбачає незалежне навчання декількох моделей на різних підвибірках даних із подальшим об'єднанням результатів прогнозування. Використання *bagging* дозволяє зменшити варіативність прогнозів та підвищити стійкість моделей до перенавчання.

До таких методів належить *Random Forest*, що базується на використанні ансамблю дерев рішень [25]. Підсумковий результат формується шляхом агрегування прогнозів окремих моделей (рисунок 2.5), що дозволяє підвищити стійкість алгоритму до шуму в даних та покращити здатність до узагальнення.

У задачах спортивної аналітики *Random Forest* використовується для прогнозування результатів матчів, оцінювання ефективності атакувальних дій та аналізу факторів, що впливають на результативність команд.

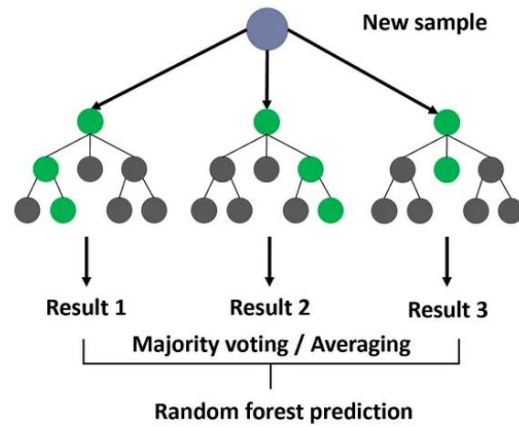
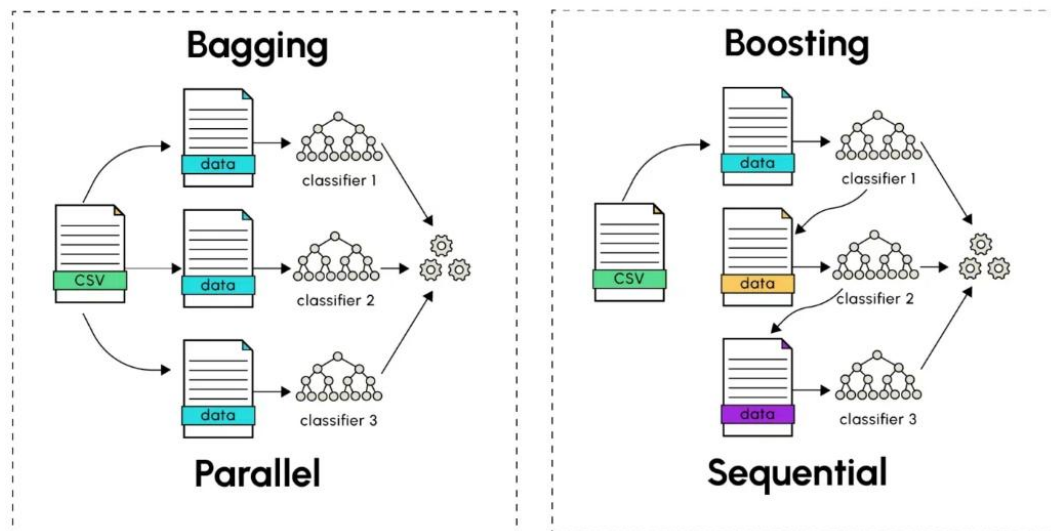


Рисунок 2.5 – Random forest [25]

Іншим важливим підходом є *boosting*, у якому моделі навчаються послідовно, а кожна наступна модель враховує помилки попередніх. Порівняння принципів роботи ансамблевих методів *bagging* та *boosting* наведено на рисунку 2.6.

Рисунок 2.6 – Порівняння підходів *bagging* та *boosting* в ансамблевих методах машинного навчання [26]

Серед сучасних алгоритмів бустингу широкого поширення набув XGBoost, що поєднує градієнтний бустинг дерев рішень із механізмами

регуляризації та оптимізації процесу навчання. Особливістю алгоритму є послідовне навчання набору дерев, де кожне наступне дерево намагається компенсувати помилки попередніх моделей. Це дозволяє ефективно виявляти складні нелінійні залежності у даних та підвищувати точність прогнозування. Завдяки високій ефективності роботи з багатовимірними ознаками XGBoost активно використовується у задачах спортивної аналітики, зокрема для прогнозування результативних дій та побудови моделей expected goals.

Для роботи з великими наборами даних також застосовується алгоритм LightGBM, орієнтований на прискорення процесу навчання та зменшення обчислювальних витрат. На відміну від XGBoost, у LightGBM дерева будуються за leaf-wise підходом, при якому розширюється вузол із найбільшою помилкою незалежно від рівня дерева (рисунок 2.7). Такий механізм дозволяє швидше зменшувати функцію втрат і підвищувати ефективність навчання моделей. Крім того, LightGBM використовує оптимізовані методи обробки даних, що забезпечує високу швидкість навчання при збереженні високої точності прогнозування [27].



Рисунок 2.7 – Порівняння принципу роботи алгоритмів XGBoost та LightGBM [27]

Використання методів машинного навчання у спортивній аналітиці забезпечує можливість виявлення складних залежностей між характеристиками ігрових епізодів та підвищує точність прогнозування результативних дій. Такі моделі здатні ефективно працювати з великою кількістю статистичних, просторових і часових ознак, автоматично визначати найбільш значущі параметри та адаптуватися до структури даних у процесі навчання.

Водночас методи машинного навчання потребують значного обсягу якісних даних та ретельної підготовки ознак. При роботі з високорозмірними спортивними даними також виникає ризик перенавчання моделей, що може призводити до погіршення якості прогнозування на нових матчах або ігрових епізодах. Крім того, окремі алгоритми мають високу обчислювальну складність та обмежену інтерпретованість результатів. У зв'язку з цим важливим етапом побудови прогнозних систем є не лише вибір алгоритму машинного навчання, а й правильне представлення та обробка футбольних даних.

2.3 Методи представлення та обробки футбольних даних

Футбольні дані характеризуються великою кількістю взаємопов'язаних просторових, часових та контекстних параметрів, що описують розвиток ігрових епізодів. Для використання таких даних у моделях машинного навчання необхідним є їх структуроване представлення у вигляді набору ознак, придатних для подальшої математичної обробки та навчання алгоритмів.

Однією з основних особливостей футбольних даних є поєднання статичних і динамічних характеристик. До статичних параметрів належать координати гравців, положення м'яча або зона виконання певної дії, тоді як динамічні характеристики описують швидкість переміщення, зміну позицій, напрямок розвитку атаки та часову послідовність подій. У зв'язку з цим

представлення футбольних даних повинно забезпечувати збереження не лише окремих статистичних показників, а й взаємозв'язків між ними у межах ігрового епізоду.

Для використання футбольних даних у задачах машинного навчання ігрові епізоди повинні бути представлені у структурованому вигляді, придатному для математичної обробки алгоритмами прогнозування. Такий підхід забезпечує можливість подальшого аналізу просторових, часових та контекстних характеристик гри, а також формування інформативних ознак для моделей машинного навчання.

Методи формування ознак (feature engineering) є одним із ключових етапів побудови моделей машинного навчання у спортивній аналітиці. Основною задачею цього процесу є перетворення сирих футбольних даних у набір інформативних параметрів, що дозволяють моделі ефективно виявляти закономірності між характеристиками ігрових епізодів та результативністю дій. Загальний процес формування ознак для моделей машинного навчання у футбольній аналітиці наведено на рисунку 2.8.

У задачах прогнозування результативних дій як ознаки можуть використовуватися просторові характеристики епізоду, зокрема відстань до воріт, кут удару, положення гравців на полі та щільність розташування захисників поблизу м'яча. Такі параметри дозволяють оцінювати рівень небезпечності атакуючої ситуації та враховувати особливості розташування гравців під час розвитку атаки.

Важливу роль також відіграють часові та контекстні ознаки, що описують динаміку розвитку ігрового епізоду. До таких параметрів належать швидкість атаки, кількість передач перед ударом, тривалість володіння м'ячем, тип передачі та характер попередніх дій команди. Використання контекстних характеристик дозволяє моделям враховувати послідовність розвитку атаки та взаємозв'язки між окремими подіями матчу [28].

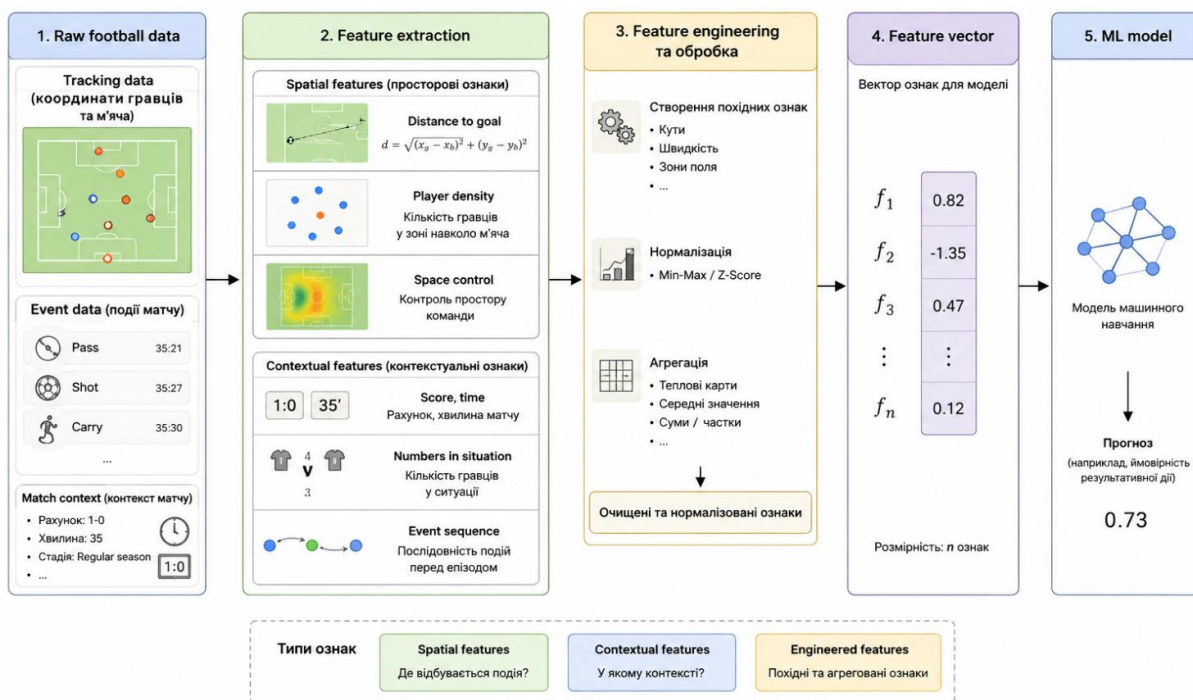


Рисунок 2.8 – Формування ознак для моделей машинного навчання у футбольній аналітиці

Для підвищення ефективності моделей застосовуються методи нормалізації та агрегування ознак [29]. Нормалізація використовується для приведення параметрів до єдиного масштабу, що забезпечує стабільність навчання алгоритмів машинного навчання. Агрегування дозволяє формувати узагальнені характеристики ігрових епізодів на основі набору окремих дій або часових інтервалів. Якість сформованих ознак безпосередньо впливає на здатність моделі виявляти складні закономірності у футбольних даних та точність прогнозування результативних дій.

Просторово-часове представлення футбольних епізодів використовується для збереження інформації про розташування гравців, переміщення м'яча та послідовність розвитку ігрових подій у межах певного проміжку часу (рисунок 2.9). На відміну від класичних статистичних показників, такий підхід дозволяє враховувати динаміку розвитку атаки, взаємодію між гравцями та зміну структури розташування команди під час матчу.

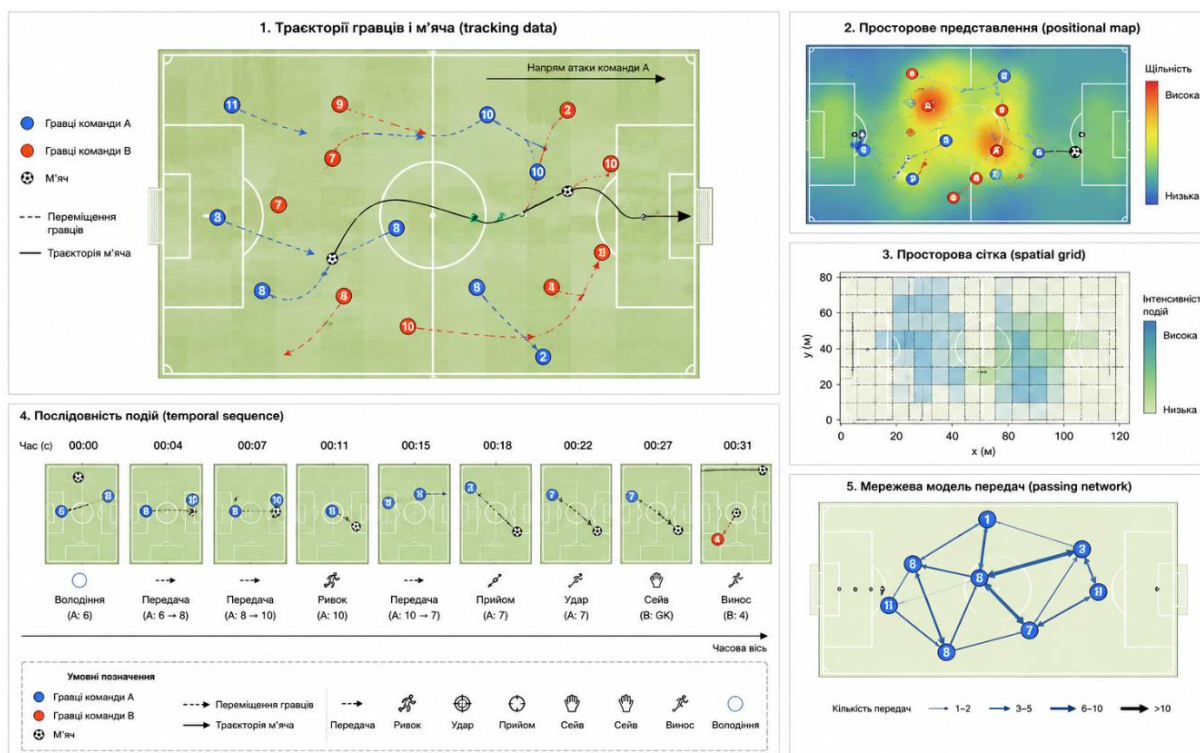


Рисунок 2.9 – Просторово-часове представлення ігрових епізодів у футбольній аналітиці

Одним із найбільш поширених способів представлення футбольних даних є координатна модель, у якій кожна подія або положення гравця описуються координатами на полі. Використання координатного представлення дозволяє аналізувати траєкторії переміщення гравців і м'яча, оцінювати просторові взаємозв'язки між учасниками епізоду та визначати найбільш небезпечні зони поля.

Для врахування часової структури гри застосовуються методи представлення послідовностей подій, у яких ігровий епізод розглядається як впорядкований набір дій команди протягом певного часового інтервалу. Такий підхід дозволяє аналізувати розвиток атак, швидкість переходу між фазами гри та взаємозв'язки між окремими подіями матчу.

У сучасних системах спортивної аналітики футбольні дані також можуть представлятися у вигляді матриць, просторових сіток або графових структур [30]. Подібні підходи дозволяють ефективно враховувати

просторово-часові залежності між гравцями та використовуються у задачах прогнозування результативних дій, аналізу тактичних схем і побудови моделей машинного навчання для оцінювання ігрових епізодів.

Методи попередньої обробки футбольних даних використовуються для підготовки інформації до подальшого навчання моделей машинного навчання. Якість попередньої обробки безпосередньо впливає на стабільність роботи алгоритмів, точність прогнозування та здатність моделей коректно виявляти закономірності у даних.

Одним із важливих етапів є обробка пропущених або некоректних значень, що можуть виникати внаслідок помилок збору даних або втрати частини інформації під час фіксації ігрових подій. Для розв'язання таких проблем застосовуються методи заповнення пропущених значень, фільтрації шуму та синхронізації часових даних між різними джерелами інформації.

Для забезпечення коректної роботи алгоритмів машинного навчання також застосовуються методи нормалізації та масштабування ознак. Нормалізація дозволяє привести параметри до єдиного діапазону значень, що є особливо важливим при роботі з просторовими та часовими характеристиками, які можуть мати різні масштаби. Масштабування ознак сприяє стабільності процесу навчання та зменшує вплив окремих параметрів на роботу моделі.

Окрему роль у попередній обробці футбольних даних відіграє кодування категоріальних ознак, до яких належать типи подій, позиції гравців, типи передач або характеристики ігрових дій. Для представлення таких параметрів у числовому вигляді використовуються методи label encoding або one-hot encoding, що дозволяє застосовувати категоріальні характеристики у моделях машинного навчання [31].

У задачах прогнозування результативних дій також важливим є балансування даних, оскільки кількість нерезультативних епізодів у футбольних матчах значно перевищує кількість голевих ситуацій.

Наявність незбалансованих даних може призводити до погіршення якості прогнозування та зміщення моделі у бік домінуючого класу. Для розв'язання цієї проблеми використовуються методи oversampling, undersampling та балансування ваг класів під час навчання моделей.

Сучасні підходи до представлення та обробки футбольних даних забезпечують можливість детального аналізу просторово-часових характеристик ігрових епізодів та підвищують ефективність моделей машинного навчання у задачах прогнозування результативних дій. Використання методів формування ознак, просторово-часового представлення та попередньої обробки даних дозволяє враховувати складні взаємозв'язки між подіями матчу та покращувати якість прогнозування.

Водночас ефективність моделей значною мірою залежить від якості підготовки даних, коректності сформованих ознак та обсягу доступної інформації. У зв'язку з цим важливим етапом побудови інтелектуальних систем спортивної аналітики є не лише вибір алгоритму машинного навчання, а й оцінювання якості отриманих прогнозів, що розглядається у наступному підрозділі.

2.4 Методи оцінювання якості моделей прогнозування

Оцінювання якості моделей є одним із ключових етапів побудови систем прогнозування у спортивній аналітиці. Основною задачею цього процесу є визначення здатності моделі коректно прогнозувати результати на нових даних та оцінювання ефективності використаних алгоритмів машинного навчання. Вибір методів оцінювання залежить від типу задачі, структури даних та особливостей прогнозованих подій.

У задачах прогнозування результативних дій у футболі особливу складність становить незбалансованість даних, оскільки кількість голевих епізодів є значно меншою порівняно з кількістю нерезультативних дій. У таких умовах використання лише загальної точності класифікації може

призводити до некоректного оцінювання ефективності моделей. Наприклад, модель може демонструвати високе значення асигасу навіть у випадку, коли більшість результативних епізодів прогнозуються неправильно.

Для оцінювання моделей прогнозування зазвичай використовується матриця помилок (confusion matrix), що відображає співвідношення правильних і помилкових прогнозів моделі [32]. Матриця (рисунок 2.10) містить чотири основні компоненти: True Positive (TP), True Negative (TN), False Positive (FP) та False Negative (FN). На основі цих значень обчислюються основні метрики якості класифікації, що дозволяють комплексно оцінити ефективність роботи моделі.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Рисунок 2.10 – Матриця помилок [32]

До основних метрик оцінювання моделей класифікації належать асигасу, precision, recall та F1-score. Ці показники обчислюються на основі значень матриці помилок та дозволяють оцінити різні аспекти якості прогнозування.

Метрика асигасу характеризує загальну частку правильно класифікованих об'єктів та визначається як відношення кількості правильних прогнозів до загальної кількості прикладів:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (2.1)$$

Незважаючи на простоту інтерпретації, асигасу може бути недостатньо інформативною у випадку незбалансованих даних, коли один із класів суттєво переважає інший.

Для оцінювання якості позитивних прогнозів використовується метрика *precision*, що визначає частку правильно передбачених позитивних подій серед усіх позитивних прогнозів моделі:

$$Precision = \frac{TP}{TP+FP}. \quad (2.2)$$

Метрика *recall* характеризує здатність моделі виявляти позитивні приклади та визначається як частка правильно знайдених позитивних об'єктів серед усіх реальних позитивних випадків:

$$Recall = \frac{TP}{TP+FN}. \quad (2.3)$$

У задачах прогнозування результативних дій *recall* є важливою характеристикою, оскільки дозволяє оцінити здатність моделі виявляти потенційно небезпечні ігрові епізоди.

Для комплексного оцінювання якості класифікації використовується метрика *F1-score*, що є гармонійним середнім між *precision* та *recall*:

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}. \quad (2.4)$$

Використання *F1-score* дозволяє отримати збалансовану оцінку ефективності моделі, особливо в умовах незбалансованих даних, характерних для задач прогнозування результативних дій у футболі.

Для більш детального оцінювання якості моделей класифікації у спортивній аналітиці також використовуються ROC-крива та метрика ROC-AUC. Такі методи дозволяють оцінити здатність моделі розділяти позитивні та негативні класи при різних значеннях порогу класифікації.

ROC-крива (Receiver Operating Characteristic curve) відображає залежність між часткою правильно визначених позитивних прикладів та

часткою хибнопозитивних прогнозів (рисунок 2.11). Чим ближче ROC-крива розташована до верхнього лівого кута координатної площини, тим кращою є здатність моделі правильно розділяти класи.

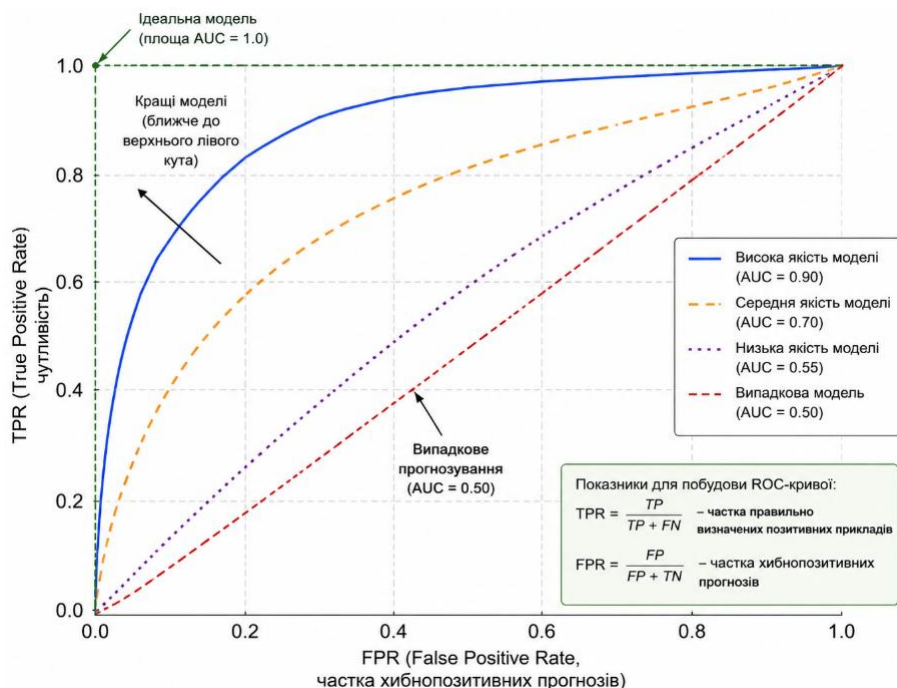


Рисунок 2.11 – ROC-крива для оцінювання якості моделей класифікації

Для кількісного оцінювання якості ROC-кривої використовується метрика ROC-AUC (Area Under Curve), що характеризує площу під ROC-кривою. Значення ROC-AUC знаходиться в діапазоні від 0 до 1, де значення, близькі до 1, свідчать про високу якість класифікації, а значення, близькі до 0.5, відповідають випадковому прогнозуванню.

У задачах спортивної аналітики метрика ROC-AUC широко використовується для оцінювання моделей прогнозування результативних дій, оскільки дозволяє враховувати якість класифікації при різних порогах прийняття рішення.

Оскільки моделі прогнозування результативних дій у футболі формують не лише класи, а й ймовірності настання подій, важливим етапом оцінювання є аналіз якості ймовірнісних прогнозів. Для цього

використовуються спеціалізовані метрики, що дозволяють оцінити ступінь відповідності між передбаченими ймовірностями та реальними результатами ігрових епізодів [33].

Однією з найбільш поширених метрик є Log Loss, або логарифмічна функція втрат, що оцінює якість прогнозованих ймовірностей з урахуванням впевненості моделі у власних прогнозах. Формула Log Loss має вигляд:

$$LogLoss = -\frac{1}{N} \sum_{i=1}^N (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)), \quad (2.5)$$

де N – кількість прикладів;

y_i – реальне значення класу;

p_i – прогнозована ймовірність позитивного класу.

Менше значення Log Loss відповідає кращій якості моделі. Особливістю цієї метрики є сильне штрафування впевнених, але помилкових прогнозів, що є важливим у задачах прогнозування результативних дій.

Для оцінювання точності ймовірнісних прогнозів також використовується метрика Brier Score, що визначає середньоквадратичну різницю між прогнозованими ймовірностями та фактичними результатами:

$$BrierScore = -\frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2. \quad (2.6)$$

Метрика Brier Score дозволяє оцінити, наскільки прогнозовані ймовірності відповідають реальній частоті виникнення подій. Нижчі значення метрики свідчать про кращу якість калібрування моделі.

Важливим аспектом оцінювання ймовірнісних моделей є calibration – відповідність між прогнозованими ймовірностями та фактичними результатами. Добре калібрована модель повинна забезпечувати ситуацію,

за якої події з прогнозованою ймовірністю 0.7 дійсно виникають приблизно у 70% випадків. Аналіз калібрування дозволяє оцінити надійність прогнозів моделі та є важливим етапом оцінювання систем прогнозування у спортивній аналітиці.

Для оцінювання ефективності моделей прогнозування доцільно використовувати комплексний підхід, що передбачає аналіз декількох метрик одночасно. Використання лише одного показника не дозволяє повною мірою оцінити якість роботи моделі, особливо в умовах незбалансованих даних, характерних для задач прогнозування результативних дій у футболі.

Комплексне використання методів оцінювання дозволяє виконувати порівняльний аналіз моделей машинного навчання, визначати їх сильні та слабкі сторони, а також обирати найбільш ефективні підходи для задач спортивної аналітики.

3 РОЗРОБКА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ПРОГНОЗУВАННЯ РЕЗУЛЬТАТИВНОСТІ УДАРІВ

Практична реалізація інтелектуальної системи прогнозування результативності ударів у футболі виконувалася мовою програмування Python із використанням середовища Google Colab. Обране середовище забезпечило можливість роботи з великими обсягами спортивних даних, проведення експериментів із моделями машинного навчання та використання хмарних обчислювальних ресурсів без необхідності локального налаштування програмного забезпечення. Крім того, Google Colab дозволив організувати процес розробки у вигляді інтерактивних ноутбуків із послідовним виконанням етапів підготовки даних, навчання моделей та аналізу результатів.

Для реалізації системи було використано набір спеціалізованих бібліотек для обробки даних, машинного навчання, спортивної аналітики та візуалізації. Завантаження та обробка футбольних event-даних здійснювалися за допомогою бібліотеки StatsBombPy, яка надає програмний доступ до відкритих даних платформи StatsBomb [34]. Для побудови футбольних візуалізацій і відображення ігрових епізодів використовувалася бібліотека mplsoccer.

Обробка та аналіз даних виконувалися із застосуванням бібліотек NumPy та Pandas, що забезпечують ефективну роботу з табличними структурами, числовими масивами та статистичними операціями. Для побудови графіків і візуального аналізу використовувалися бібліотеки Matplotlib та Seaborn.

Реалізація моделей машинного навчання виконувалася на основі бібліотеки scikit-learn, яка використовувалася для підготовки ознак, масштабування даних, поділу вибірки та навчання базових моделей класифікації. У межах дослідження було реалізовано логістичну регресію,

випадковий ліс та багатошаровий перцептрон. Для реалізації ансамблевих градієнтних моделей використовувалися бібліотеки XGBoost та CatBoost.

Оцінювання якості моделей здійснювалося із використанням метрик accuracy, precision, recall, F1-score, ROC-AUC, log-loss та Brier Score. Використання метрики Brier Score дозволило додатково оцінити якість прогнозованих імовірностей та ступінь калібрування моделей. Додатково застосовувалися матриці помилок, ROC-криві та precision-recall-криві, що дозволило провести комплексний аналіз якості прогнозування результативних ударів.

Користувацький інтерфейс системи реалізовано за допомогою фреймворку Streamlit. Це дозволило створити інтерактивний вебзастосунок для введення параметрів удару, візуалізації положення гравця на полі та отримання прогнозованої ймовірності результативної дії у режимі реального часу.

3.1 Формування та підготовка датасету футбольних ударів

3.1.1 Формування набору футбольних матчів та event-даних

Формування датасету для навчання моделей прогнозування результативності ударів розпочиналося із завантаження повного списку доступних футбольних турнірів із відкритого набору даних StatsBomb. Отриманий список (рисунк 3.1) містив інформацію про турніри, сезони та відповідні ідентифікатори, необхідні для подальшого доступу до матчів і event-даних.

Після отримання загального переліку турнірів було виконано фільтрацію з метою формування репрезентативного набору сучасних чоловічих топових футбольних змагань. Для кожного турніру додатково перевірялася доступність матчів та кількість доступних ігрових подій у межах окремих сезонів.

	competition_id	season_id	competition_name	season_name
0	9	27	1. Bundesliga	2015/2016
1	9	281	1. Bundesliga	2023/2024
2	1267	107	African Cup of Nations	2023
3	16	276	Champions League	1970/1971
4	16	71	Champions League	1971/1972
...	...	...	...	...
70	35	75	UEFA Europa League	1988/1989
71	53	106	UEFA Women's Euro	2022
72	53	315	UEFA Women's Euro	2025
73	72	30	Women's World Cup	2019
74	72	107	Women's World Cup	2023

Рисунок 3.1 – Списку доступних турнірів StatsBomb Open Data

На наступному етапі було проаналізовано кількість матчів для кожного турніру та сезону, що дозволило оцінити обсяг доступних даних і сформувати фінальний перелік змагань для подальшого дослідження. Турніри із найбільшою кількістю матчів наведено на рисунку 3.2.

	competition_name	season_name	matches_count
0	La Liga	2015/2016	380
1	Premier League	2015/2016	380
2	1. Bundesliga	2015/2016	306
3	FIFA World Cup	2022	64
4	FIFA World Cup	2018	64
5	UEFA Euro	2024	51
6	UEFA Euro	2020	51
7	La Liga	2014/2015	38
8	Premier League	2003/2004	38
9	La Liga	2011/2012	37

Рисунок 3.2 – Турніри з найбільшою кількістю доступних матчів

У результаті було сформовано фінальний набір турнірів, до якого увійшли матчі чемпіонату Іспанії La Liga, Англійської Premier League, німецької Bundesliga, чемпіонатів світу FIFA World Cup 2018 та 2022 років, а також турнірів UEFA Euro 2020 і UEFA Euro 2024. Використання декількох топових турнірів дозволило отримати більш різноманітний набір ігрових ситуацій та зменшити залежність моделей від специфіки окремого чемпіонату.

Після формування фінального списку турнірів було виконано автоматизоване завантаження інформації про всі доступні матчі для кожного сезону. На основі отриманих даних сформовано єдиний список ідентифікаторів матчів `match_id`, який використовувався для подальшого завантаження event-даних.

У результаті обробки було сформовано набір із 1296 матчів, для яких надалі завантажувалися детальні event-дані, включаючи передачі, удари, втрати м'яча, єдиноборства та інші ігрові події.

На завершальному етапі виконувалося послідовне завантаження event-даних для кожного матчу із подальшим об'єднанням усіх подій у єдиний датасет. У результаті було сформовано набір даних розміром 4522054 події та 121 ознака. Отримані сирі event-дані було збережено окремо для забезпечення можливості повторного використання без необхідності повторного завантаження з зовнішнього джерела.

3.1.2 Формування датасету ударів

Після завантаження повного набору event-даних було виконано виділення подій типу Shot, що відповідають ударам по воротах. Для цього з загального набору подій відбиралися лише записи, у яких поле `type` мало значення Shot. У результаті було сформовано окремий датасет ударів `shots_df`, який містив 32736 ігрових епізодів та 121 ознаку.

Наступним етапом було проведено аналіз результатів ударів. Розподіл значень поля `shot_outcome` (рисунк 3.3) показав, що найбільшу частину становили неточні удари, заблоковані удари та удари, які були відбиті воротарем. Кількість результативних ударів типу `Goal` склала 3498 епізодів.

count	
shot_outcome	
Off T	10463
Blocked	8340
Saved	7672
Goal	3498
Wayward	1896
Post	655
Saved Off Target	108
Saved to Post	104

Рисунок 3.3 – Розподіл результатів ударів у датасеті

Після цього було виконано перевірку ключових полів, пов'язаних із характеристиками ударів, зокрема типу удару, частини тіла, техніки виконання, координат та значення очікуваних голів `StatsBomb xG`. Це дозволило оцінити повноту даних та підготувати ознаки для подальшого використання у моделях машинного навчання.

Задача прогнозування результативності ударів була сформульована як задача бінарної класифікації, у межах якої модель визначає ймовірність того, що конкретний удар завершиться голом. Для цього було сформовано цільову змінну `goal`, яка приймала значення 1 у випадку результативного удару та 0 для всіх інших результатів. У результаті було отримано 3498 позитивних прикладів та 29238 нерезультативних ударів, що свідчить про наявність дисбалансу класів у датасеті.

Однією з ключових характеристик удару є його просторове розташування на футбольному полі. Для цього з поля `location` було

витагнуто координати x та y , які визначають позицію гравця у момент виконання удару.

На основі координат було розраховано додаткові просторові ознаки. Зокрема, для кожного удару визначалася евклідова дистанція до центру воріт, що дозволило кількісно оцінити складність ігрового епізоду залежно від відстані до зони завершення атаки.

Додатково було розраховано кут удару відносно площини воріт. Для цього використовувалися координати удару та стандартна ширина футбольних воріт. Отримана ознака дозволяє враховувати не лише дистанцію, а й геометричну зручність позиції для завершення атаки.

На рисунку 3.4 наведено статистичний розподіл розрахованих просторових ознак.

distance_to_goal		shot_angle	
count	32736.000000	count	32736.000000
mean	18.920109	mean	23.448599
std	8.489706	std	14.575539
min	0.632456	min	0.000000
25%	12.000000	25%	13.917013
50%	18.263078	50%	18.156597
75%	25.181144	75%	29.295920
max	86.137100	max	180.000000

Рисунок 3.4 – Статистичні характеристики ознак дистанції до воріт та кута удару

На завершальному етапі було сформовано основний датасет для машинного навчання (рисунки 3.5), який містив числові та категоріальні характеристики ударів, просторові ознаки, контекст виконання удару, значення StatsBomb xG та сформовану цільову змінну goal. Підсумковий датасет містив 32736 записів та 13 ознак.

	x	y	distance_to_goal	shot_angle	minute	period	under_pressure	shot_body_part	shot_type	shot_technique	play_pattern	shot_statsbomb_xg	goal
2356	112.5	42.1	7.788453	49.274246	2	1	1	Head	Open Play	Normal	From Corner	0.100429	0
2357	90.9	39.9	29.100172	14.337098	8	1	0	Right Foot	Open Play	Normal	From Corner	0.023138	0
2358	103.1	52.5	21.020466	16.104704	9	1	0	Right Foot	Open Play	Half Volley	From Throw In	0.034535	0
2359	116.6	38.6	3.676955	89.713616	9	1	1	Head	Open Play	Normal	From Corner	0.394774	1
2360	112.4	35.8	8.683317	41.899301	15	1	1	Head	Open Play	Normal	From Throw In	0.077194	0

Рисунок 3.5 – Фрагмент сформованого датасету

3.1.3 Дослідницький аналіз даних

Після формування основного датасету було виконано дослідницький аналіз даних з метою виявлення закономірностей між характеристиками ударів та їх результативністю. На першому етапі проаналізовано розподіл цільової змінної, а також базових геометричних ознак ударів – дистанції до воріт та кута удару. Результати аналізу наведено на рисунку 3.6.

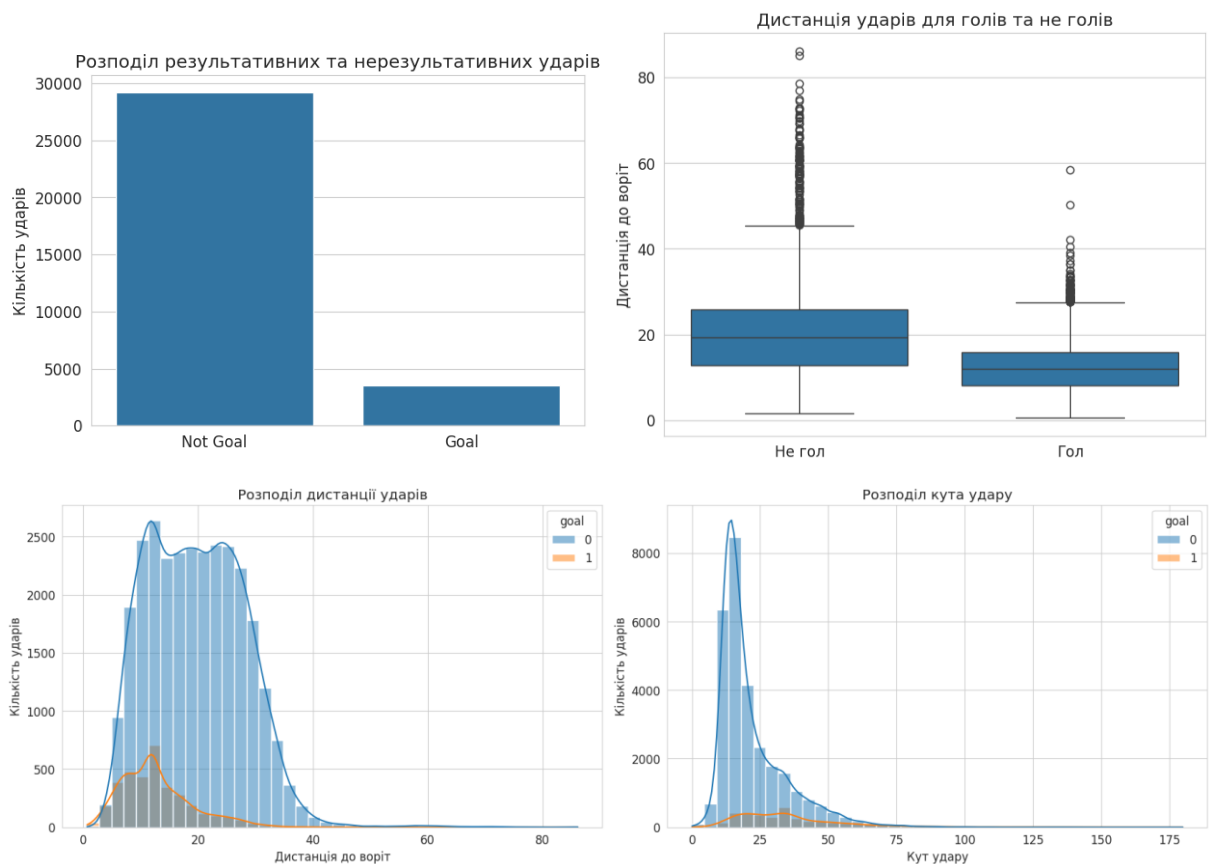


Рисунок 3.6 – Розподіл цільової змінної та геометричних характеристик ударів

Аналіз розподілу цільової змінної показав наявність вираженого дисбалансу класів у датасеті: кількість нерезультативних ударів суттєво перевищує кількість голів. Така особливість є типовою для задач прогнозування результативності ударів у футболі, оскільки більшість ударів у реальних матчах не завершуються взяттям воріт.

Розподіл дистанції ударів демонструє, що найбільша кількість ударів виконується з відстані приблизно 10–30 метрів до воріт. Водночас результативні удари значно частіше виконуються з ближчих позицій. Це також підтверджується boxplot-діаграмою, на якій медіанна дистанція для голів є помітно меншою порівняно з нерезультативними ударами.

Аналіз кута удару показав, що найбільша частка ударів виконується з відносно невеликих кутів, тоді як зі збільшенням кута ймовірність результативного завершення атаки зростає.

Для аналізу просторових закономірностей виконання ударів було побудовано візуалізації розташування ударів на полі (рисунок 3.7).

Аналіз просторового розподілу показав, що найбільша концентрація ударів спостерігається у центральній частині штрафного майданчика та поблизу зони перед воротами. Саме ці області характеризуються найвищою щільністю результативних ударів.

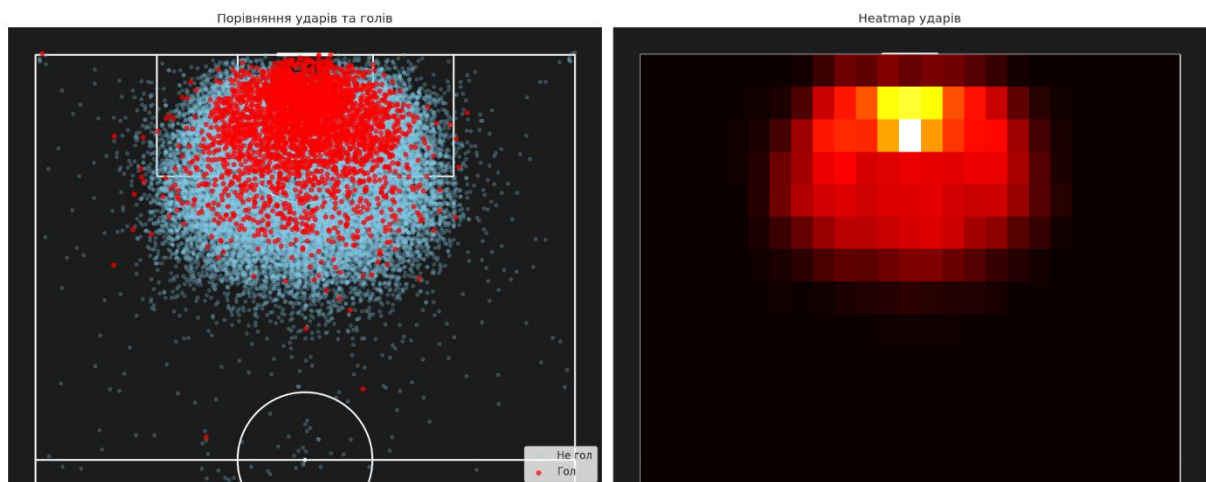


Рисунок 3.7 – Просторовий розподіл ударів на футбольному полі

Теплова карта ударів додатково підтверджує, що переважна більшість атак завершується ударами з центральних зон поблизу воріт, тоді як удари з далеких дистанцій та гострих кутів трапляються значно рідше. Отримані результати узгоджуються з виявленими раніше закономірностями щодо впливу дистанції та кута удару на ймовірність взяття воріт.

Наступним етапом дослідницького аналізу стало вивчення категоріальних характеристик ударів та їх впливу на результативність завершення атак. Для цього було проаналізовано типи ударів, частину тіла, якою виконувався удар, а також вплив тиску суперника під час завершення атаки.

Аналіз розподілу типів ударів показав, що переважна більшість ударів у датасеті належить до категорії Open Play, тобто ударів із гри без стандартних положень (рисунок 3.8). Удари зі штрафних, кутових та пенальті становлять значно меншу частину вибірки.

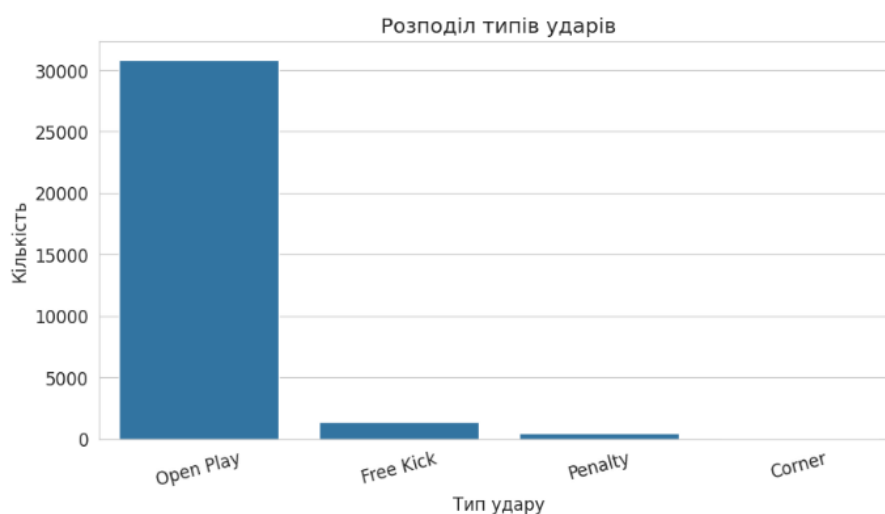


Рисунок 3.8 – Розподіл типів ударів у датасеті

Порівняння результативності різних типів ударів (рисунок 3.9) показало, що найвищу частку голів мають пенальті, що пояснюється специфікою виконання удару з мінімальною дистанцією та високою ймовірністю реалізації. Для ударів із гри (Open Play) значення

результативності є суттєво нижчим, тоді як удари зі штрафних та кутових демонструють ще меншу частку результативних завершень.

Аналіз залежності результативності від частини тіла показав близькі значення для ударів головою та ударами ногою, тоді як категорія Other демонструє вищу частку голів через невелику кількість специфічних ігрових епізодів у вибірці (рисунок 3.10).

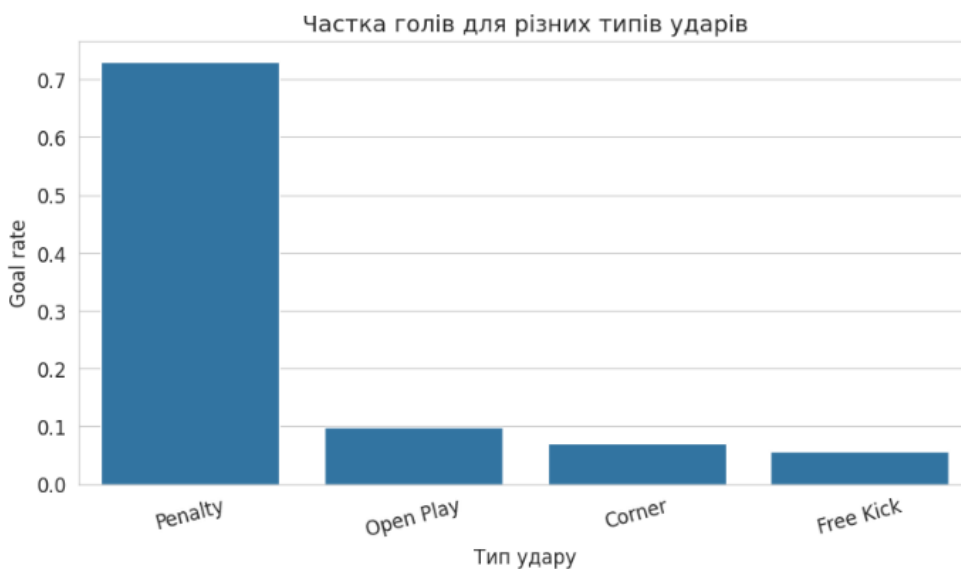


Рисунок 3.9 – Частка результативних ударів для різних типів ударів

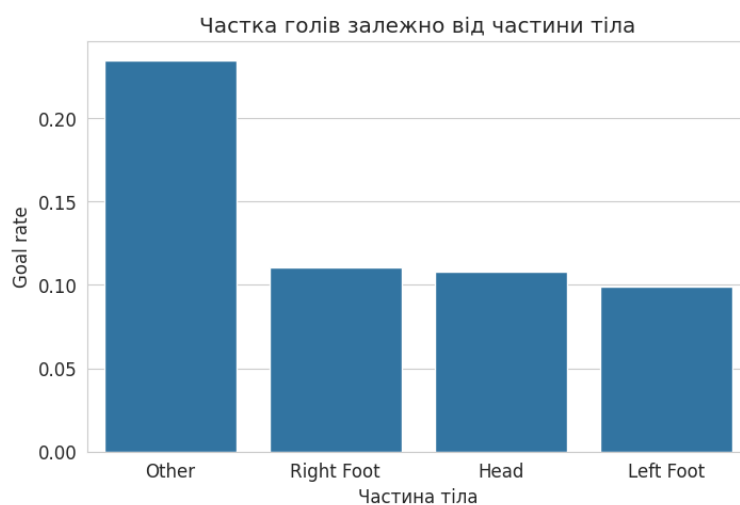


Рисунок 3.10 – Частка результативних ударів залежно від частини тіла

Додатково було проаналізовано вплив тиску суперника на результативність ударів. Отримані результати показали, що удари без активного тиску з боку захисників мають вищу ймовірність завершення голом порівняно з ударами, виконаними під тиском. Це підтверджує важливість контекстних ігрових характеристик для задач прогнозування результативності ударів.

На завершальному етапі дослідницького аналізу було проаналізовано взаємозв'язки між числовими ознаками датасету та залежність між просторовими характеристиками ударів і значенням показника StatsBomb xG.

Аналіз кореляційної матриці (рисунок 3.11) показав наявність сильного від'ємного зв'язку між координатою x та дистанцією до воріт, що пояснюється особливостями системи координат футбольного поля. Також спостерігається виражена від'ємна кореляція між дистанцією до воріт та кутом удару: зі зменшенням дистанції кут атаки зазвичай стає більш сприятливим для завершення епізоду.

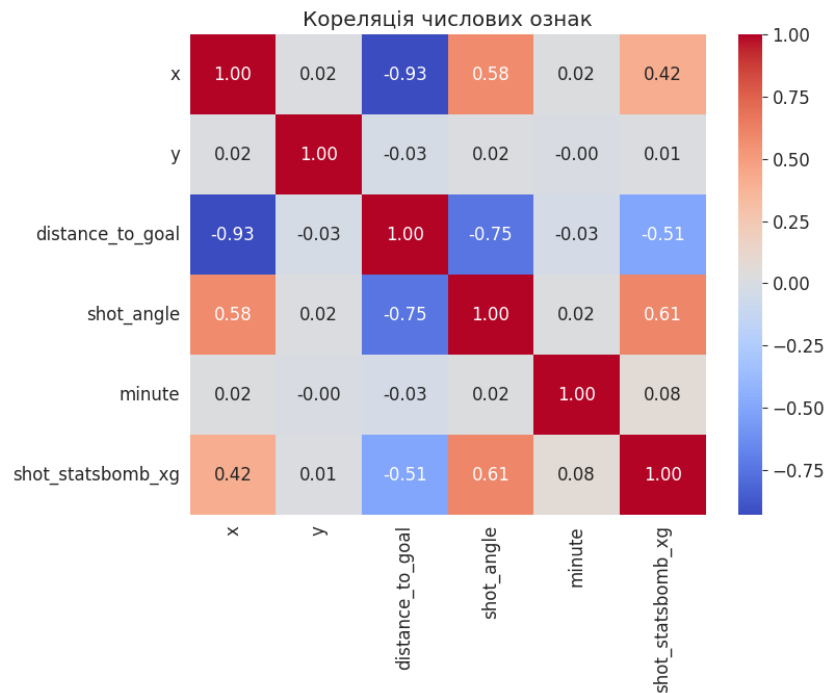


Рисунок 3.11 – Кореляційна матриця числових ознак датасету ударів

Аналіз кореляційної матриці показав сильний від'ємний зв'язок між дистанцією до воріт та кутом удару, а також помірну кореляцію між просторовими ознаками та значенням StatsBomb xG. Для детальнішого аналізу цього зв'язку було побудовано залежності між показником StatsBomb xG та просторовими характеристиками ударів (рисунок 3.12).

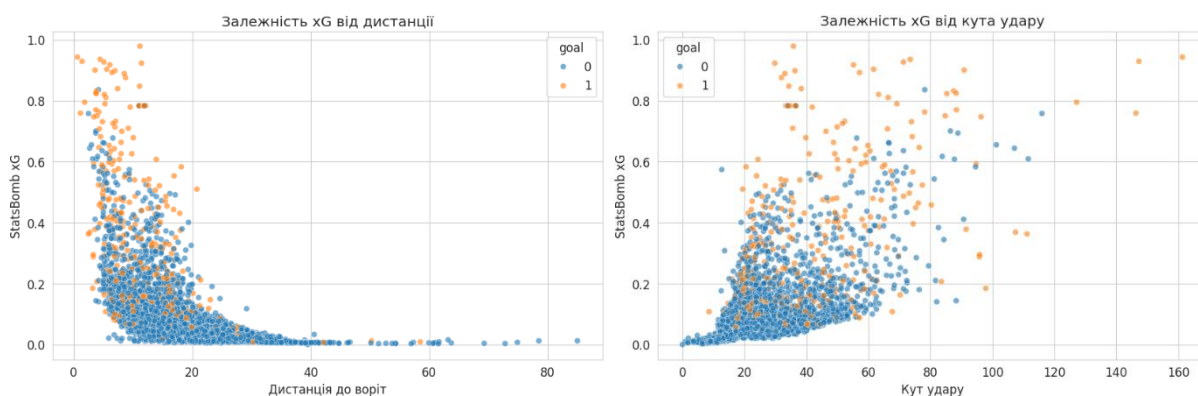


Рисунок 3.12 – Залежність показника StatsBomb xG від геометричних характеристик удару

Побудовані залежності показали, що найбільші значення xG характерні для ударів із близької дистанції та більш вигідного кута атаки. Зі збільшенням дистанції значення xG переважно зменшуються, тоді як зі збільшенням кута удару ймовірність результативного завершення атаки зростає. Порівняння сформованих ознак із професійною метрикою StatsBomb xG підтвердило, що дистанція та кут удару коректно відображають небезпечність ігрового епізоду.

3.1.4 Попередня обробка даних для моделей машинного навчання

Перед навчанням моделей було виконано попередню обробку ознак, необхідну для коректного використання числових і категоріальних даних у алгоритмах машинного навчання. Спочатку було згруповано рідкісні категорії, які мали малу кількість спостережень у датасеті. Зокрема, окремі

малочисельні значення у полях `shot_type`, `shot_technique` та `shot_body_part` було об'єднано в узагальнену категорію `Rare` або `Other`. Це дозволило зменшити кількість слабо представлених категорій та уникнути надмірного розрідження ознак після кодування.

Для навчання моделей було сформовано матрицю ознак X та цільовий вектор y . До числових ознак увійшли координати удару, дистанція до воріт, кут удару, хвилина матчу, період гри, значення `StatsBomb xG` та ознака тиску суперника. До категоріальних ознак було віднесено частину тіла, тип удару, техніку виконання та ігровий патерн. Загальний розмір матриці ознак становив 32736 записів та 12 вхідних характеристик.

Після формування ознак датасет було поділено на навчальну та тестову вибірки у співвідношенні 80:20. Поділ виконувався зі стратифікацією за цільовою змінною, що дозволило зберегти початкове співвідношення результативних і нерезультативних ударів в обох частинах вибірки. У результаті навчальна вибірка містила 26188 записів, а тестова – 6548 записів.

Для уніфікованої підготовки даних було використано `ColumnTransformer`, який окремо обробляв числові та категоріальні ознаки (рисунок 3.13). Числові характеристики масштабувалися за допомогою `StandardScaler`, а категоріальні ознаки перетворювалися за допомогою `OneHotEncoder` з параметром `handle_unknown='ignore'`, що дозволило коректно обробляти категорії, які могли бути відсутніми під час навчання, але з'явитися у тестових або користувацьких даних.

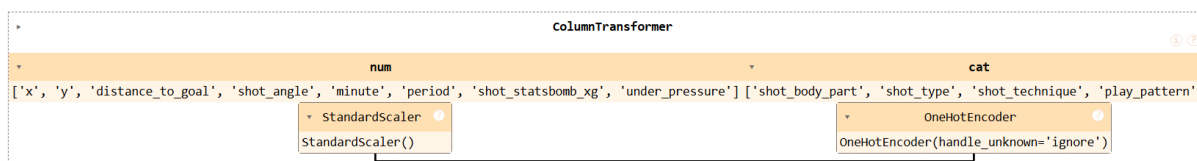


Рисунок 3.13 – Структура препроцесора для числових і категоріальних ознак

3.2 Реалізація моделей машинного навчання

Після завершення етапу підготовки даних було реалізовано набір моделей машинного навчання для прогнозування результативності ударів у футболі. У межах дослідження використовувалися як базові алгоритми бінарної класифікації, так і ансамблеві методи та нейромережеві підходи. Зокрема, було реалізовано моделі Logistic Regression, Random Forest, XGBoost, CatBoost та багатошаровий перцептрон (MLP).

Усі моделі реалізовувалися із використанням Pipeline бібліотеки scikit-learn, що дозволило об'єднати preprocessing ознак та класифікатор у єдину послідовність обробки даних. Такий підхід забезпечив однакоvu схему підготовки ознак для всіх моделей та спростив процес навчання й подальшого прогнозування.

Навчання моделей виконувалося на тренувальній вибірці, сформованій на попередньому етапі, а оцінювання якості прогнозування проводилося на тестових даних. Для моделей, що підтримують імовірнісне прогнозування, додатково обчислювалися ймовірності результативності ударів, які надалі використовувалися для побудови ROC-кривих, Precision-Recall-кривих, аналізу калібрування та порівняння моделей.

3.2.1 Logistic Regression

Першою моделлю, реалізованою у межах дослідження, стала логістична регресія. Даний алгоритм було обрано як базову модель бінарної класифікації для прогнозування результативності ударів. Загальну структуру pipeline для логістичної регресії наведено на рисунку 3.14.

Під час налаштування моделі використовувався параметр `class_weight='balanced'`, оскільки датасет містив суттєвий дисбаланс між результативними та нерезультативними ударами. Це дозволило зменшити зміщення моделі у бік домінуючого класу. Параметр `max_iter=2000` було

використано для забезпечення стабільної збіжності алгоритму під час навчання на великому наборі даних.

Після навчання модель використовувалася для прогнозування класів та ймовірностей результативності ударів на тестовій вибірці.

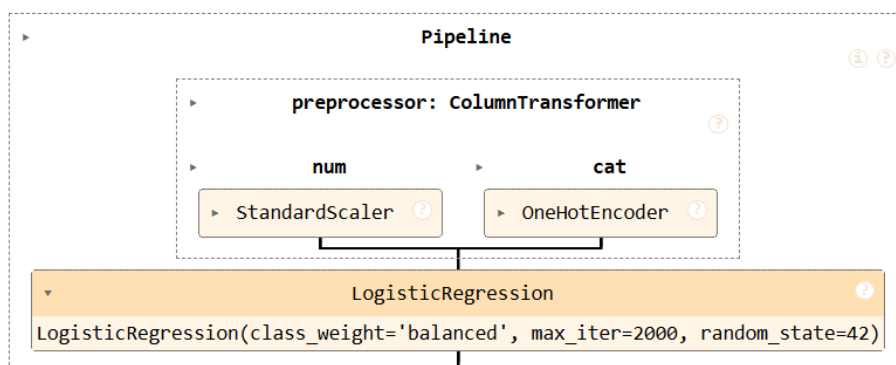


Рисунок 3.14 – Pipeline моделі Logistic Regression

3.2.2 Random Forest

Наступною моделлю, реалізованою у межах дослідження, став алгоритм випадкового лісу (Random Forest). Даний метод належить до ансамблевих алгоритмів машинного навчання та базується на побудові множини незалежних дерев рішень із подальшим агрегуванням їх прогнозів. Використання випадкового лісу дозволяє враховувати нелінійні залежності між ознаками та зменшує ризик перенавчання моделі. Загальну структуру pipeline моделі наведено на рисунку 3.15.

Під час налаштування моделі використовувався параметр `class_weight='balanced'` для компенсації дисбалансу між результативними та нерезультативними ударами. Кількість дерев у ансамблі становила `n_estimators=300`, що дозволило підвищити стабільність прогнозування. Для обмеження складності дерев використовувалися параметри `max_depth=12`, `min_samples_split=10` та `min_samples_leaf=5`, які зменшували ризик перенавчання на тренувальних даних.

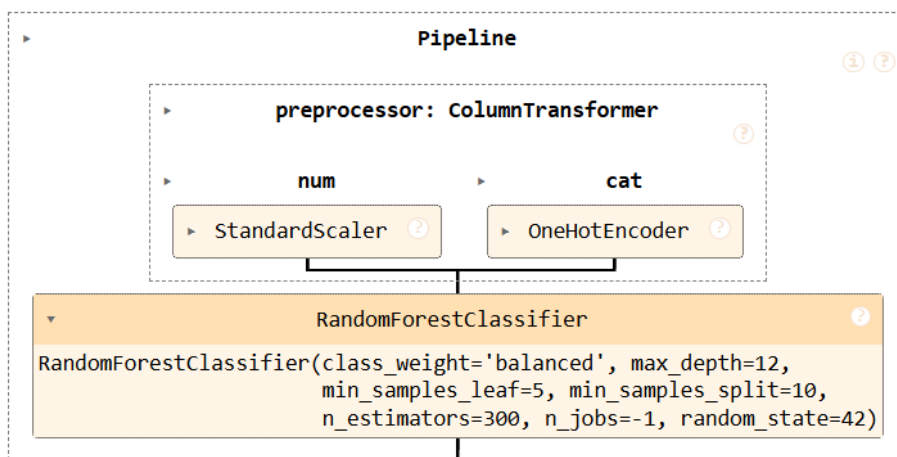


Рисунок 3.15 – Pipeline моделі Random Forest

3.2.3 XGBoost

Для подальшого підвищення якості прогнозування у роботі було реалізовано модель градієнтного бустингу XGBoost. Даний алгоритм використовує послідовне навчання ансамблю дерев рішень, у межах якого кожне наступне дерево коригує помилки попередніх моделей. Такий підхід дозволяє ефективно виявляти складні нелінійні залежності між ознаками та забезпечує високу якість класифікації для табличних даних.

Під час налаштування моделі використовувалися параметри `n_estimators=400`, `max_depth=5` та `learning_rate=0.05`, що дозволило сформувати достатньо глибокий ансамбль із поступовим навчанням дерев рішень. Для зменшення ризику перенавчання застосовувалися параметри `subsample=0.8` та `colsample_bytree=0.8`, які забезпечували випадковий відбір частини спостережень та ознак під час побудови окремих дерев. Загальну структуру pipeline моделі наведено на рисунку 3.16.

Оскільки задача прогнозування голів характеризується вираженим дисбалансом класів, додатково було виконано підбір оптимального порогу класифікації (`threshold tuning`) замість використання стандартного значення 0.5.

Для цього аналізувалися зміни метрик precision, recall та F1-score при різних значеннях порогу класифікації. Результати аналізу наведено на рисунку 3.17.

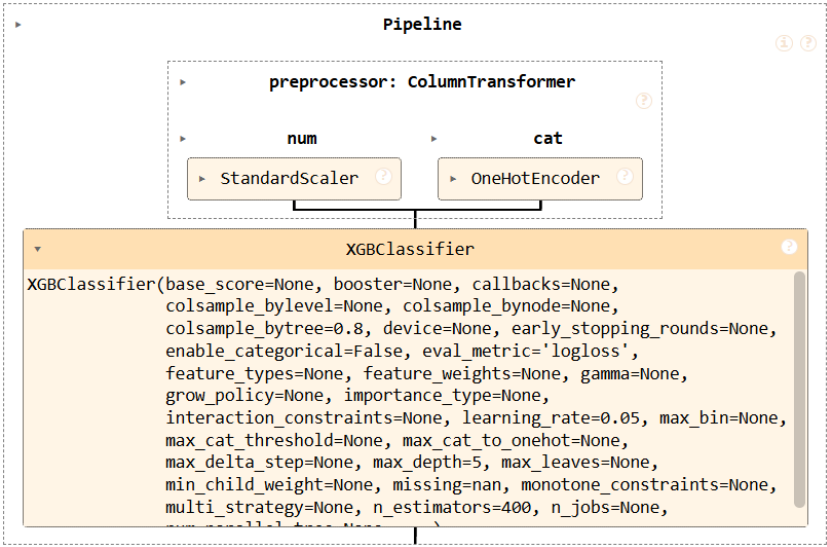


Рисунок 3.16 – Pipeline моделі XGBoost

	threshold	precision	recall	f1
0	0.10	0.275698	0.747143	0.402772
1	0.15	0.343982	0.661429	0.452590
2	0.20	0.400000	0.580000	0.473469
3	0.25	0.451372	0.517143	0.482024
4	0.30	0.507414	0.440000	0.471308
5	0.35	0.559761	0.401429	0.467554
6	0.40	0.624365	0.351429	0.449726
7	0.45	0.666667	0.311429	0.424537
8	0.50	0.700000	0.280000	0.400000
9	0.55	0.733607	0.255714	0.379237
10	0.60	0.747619	0.224286	0.345055
11	0.65	0.770115	0.191429	0.306636
12	0.70	0.785714	0.172857	0.283372
13	0.75	0.801724	0.132857	0.227941
14	0.80	0.840580	0.082857	0.150845
15	0.85	0.923077	0.051429	0.097429
16	0.90	0.913043	0.030000	0.058091

Рисунок 3.17 – Аналіз впливу порогу класифікації на метрики моделі XGBoost

Отримані результати показали, що зі збільшенням порогу класифікації значення precision поступово зростає, тоді як recall зменшується. Максимальне значення F1-score було досягнуто при порозі 0.25, який і використовувався як оптимальний для подальшого прогнозування результативних ударів. Використання зниженого порогу дозволило покращити баланс між точністю та повнотою моделі в умовах дисбалансу класів.

3.2.4 CatBoost

Ще однією моделлю, реалізованою у межах дослідження, став алгоритм градієнтного бустингу CatBoost. Дана модель орієнтована на ефективну роботу з табличними даними та забезпечує стабільне навчання ансамблю дерев рішень із використанням градієнтного бустингу. Використання CatBoost дозволяє зменшити ризик перенавчання та підвищити якість прогнозування для задач класифікації.

Під час налаштування моделі використовувалися параметри `iterations=300`, `depth=6` та `learning_rate=0.05`, які визначали кількість дерев у ансамблі, глибину дерев та швидкість навчання моделі відповідно. Як функцію втрат використовувалося `Logloss`, а основною метрикою оцінювання під час навчання було обрано `AUC`. Загальну структуру pipeline моделі наведено на рисунку 3.18.

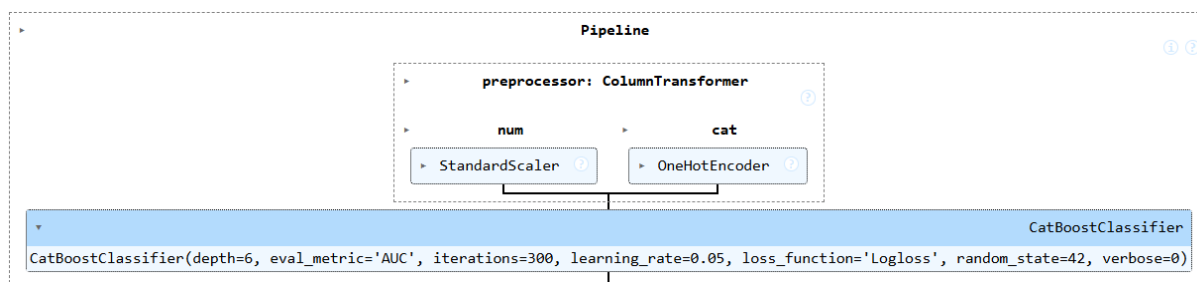


Рисунок 3.18 – Pipeline моделі CatBoost

Аналогічно до моделі XGBoost, для CatBoost було виконано підбір оптимального порогу класифікації замість використання стандартного значення 0.5. Це було необхідно через дисбаланс між результативними та нерезультативними ударами у датасеті.

Для різних значень порогу класифікації аналізувалися зміни метрик precision, recall та F1-score. Результати threshold tuning наведено на рисунку 3.19.

	threshold	precision	recall	f1
0	0.10	0.280881	0.747143	0.408275
1	0.15	0.353788	0.667143	0.462376
2	0.20	0.410696	0.581429	0.481372
3	0.25	0.469974	0.514286	0.491132
4	0.30	0.523490	0.445714	0.481481
5	0.35	0.573196	0.397143	0.469198
6	0.40	0.631854	0.345714	0.446907
7	0.45	0.691824	0.314286	0.432220
8	0.50	0.732558	0.270000	0.394572
9	0.55	0.746725	0.244286	0.368138
10	0.60	0.766667	0.230000	0.353846
11	0.65	0.788571	0.197143	0.315429
12	0.70	0.793548	0.175714	0.287719
13	0.75	0.798387	0.141429	0.240291
14	0.80	0.903846	0.067143	0.125000
15	0.85	0.918919	0.048571	0.092266
16	0.90	0.904762	0.027143	0.052705

Рисунок 3.19 – Аналіз впливу порогу класифікації на метрики моделі CatBoost

Отримані результати показали аналогічну до XGBoost поведінку метрик: зі збільшенням порогу precision зростає, тоді як recall поступово зменшується. Максимальне значення F1-score було досягнуто при порозі 0.25, який надалі використовувався як оптимальний для прогнозування результативних ударів. Використання оптимізованого порогу дозволило покращити баланс між точністю та повнотою моделі.

3.2.5 MLP

Для дослідження можливостей нейромережевих підходів у задачі прогнозування результативності ударів було реалізовано багатошаровий перцептрон (MLPClassifier). Дана модель належить до штучних нейронних мереж та дозволяє моделювати складні нелінійні залежності між ознаками, які можуть бути недоступними для класичних лінійних алгоритмів.

Архітектура моделі містила два приховані шари розміром 64 та 32 нейрони відповідно. Наявність декількох прихованих шарів дозволяє моделі формувати більш складні внутрішні представлення ознак та враховувати комплексні взаємозв'язки між параметрами ігрових епізодів. Як функцію активації використовувалася ReLU, а оптимізація параметрів виконувалася за допомогою алгоритму Adam. Загальну структуру pipeline моделі наведено на рисунку 3.20.

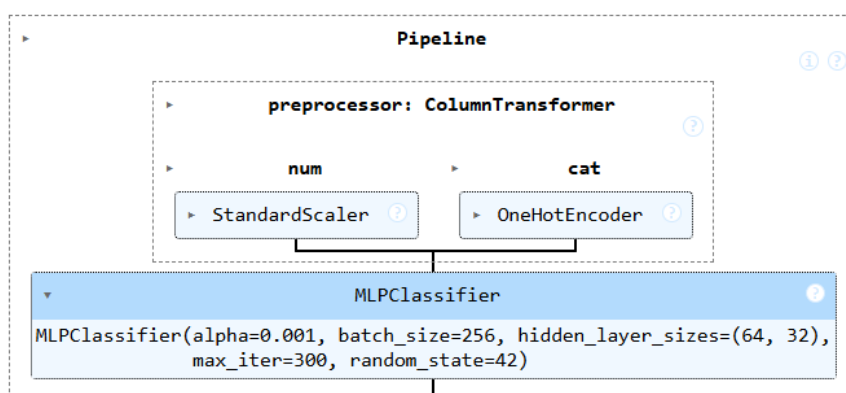


Рисунок 3.20 – Pipeline моделі MLP

Під час налаштування моделі використовувалися параметри `alpha=0.001` для регуляризації ваг нейронної мережі та `batch_size=256` для пакетного навчання. Максимальна кількість епох навчання становила `max_iter=300`. Використання багатошарового перцептрона дозволило дослідити здатність нейронної мережі виявляти складні взаємозв'язки між просторовими, часовими та контекстними характеристиками ударів.

Після завершення навчання та налаштування моделей було виконано збереження навченої моделі разом із preprocessing-блоком для подальшого використання у системі. Використання Pipeline дозволило зберегти єдину послідовність обробки даних та прогнозування, що забезпечило коректну інтеграцію моделі у користувацький інтерфейс системи прогнозування результативності ударів.

3.3 Розробка користувацького інтерфейсу системи

Для забезпечення практичного використання розробленої моделі прогнозування було реалізовано інтерактивний користувацький інтерфейс на основі фреймворку Streamlit. Використання Streamlit дозволило створити програмний інтерфейс із можливістю введення параметрів удару, автоматичного формування прогнозу та візуалізації результатів роботи моделі машинного навчання у режимі реального часу.

Розроблений інтерфейс забезпечує взаємодію користувача із системою без необхідності роботи з програмним кодом або середовищем розробки. Після введення або зміни параметрів удару система автоматично виконує попередню обробку даних, формує вектор ознак та передає його до навченої моделі машинного навчання для прогнозування ймовірності голу.

Загальна схема роботи системи прогнозування результативності ударів наведена на рисунку 3.21. На першому етапі користувач через вебінтерфейс вводить параметри ігрової ситуації, включаючи координати удару, тип удару, техніку виконання, частину тіла, хвилину матчу та інші характеристики епізоду. Після цього система виконує попередню обробку даних, що включає масштабування числових ознак, кодування категоріальних параметрів та формування вхідного вектора ознак відповідно до структури навченої моделі.

Сформований набір ознак передається до збереженої моделі машинного навчання, яка здійснює прогнозування ймовірності голу за

допомогою методу `predict_proba()`. Отриманий результат використовується для формування текстового прогнозу та оцінювання рівня небезпеки удару. Додатково система виконує графічне відображення положення удару на футбольному полі та візуалізацію напрямку завершення атаки.



Рисунок 3.21 – Загальна схема роботи системи

Структурно інтерфейс складається з бічної панелі навігації та основної робочої області (рисунок 3.22). У бічній панелі реалізовано перемикання між двома функціональними модулями системи: модулем аналізу реального удару та модулем моделювання ігрової ситуації.

Першим функціональним модулем системи є модуль аналізу реального удару, призначений для дослідження ігрових епізодів із підготовленого датасету футбольних ударів (рисунок 3.23). У межах цього модуля користувач може обрати конкретний удар із датасету за допомогою інтерактивного елемента керування, після чого система автоматично завантажує відповідні параметри ігрової ситуації та виконує прогнозування результативності удару.

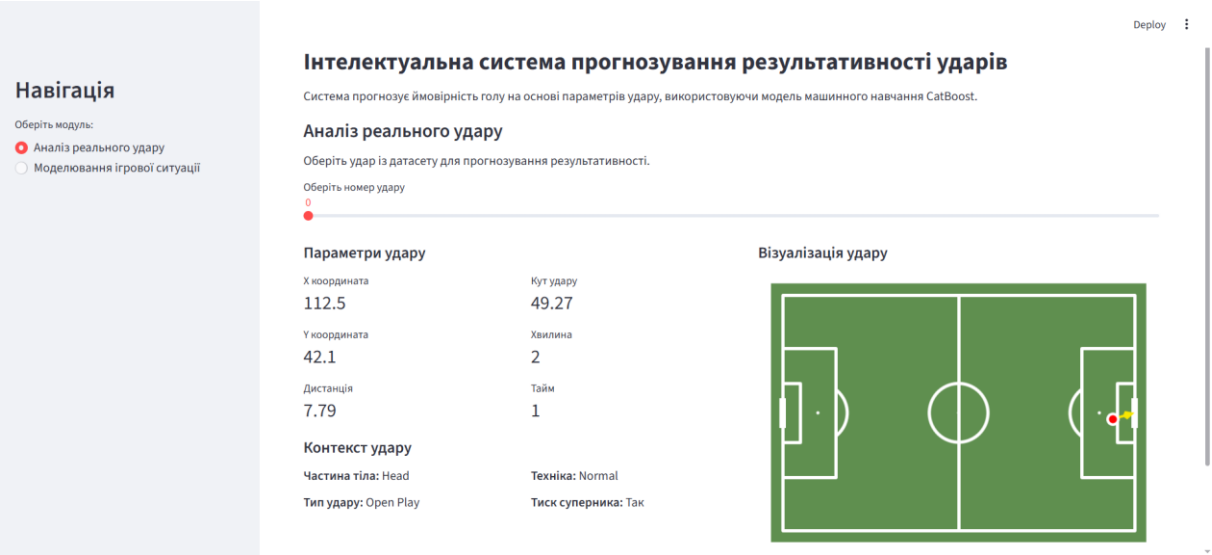


Рисунок 3.22 – Інтерфейс системи

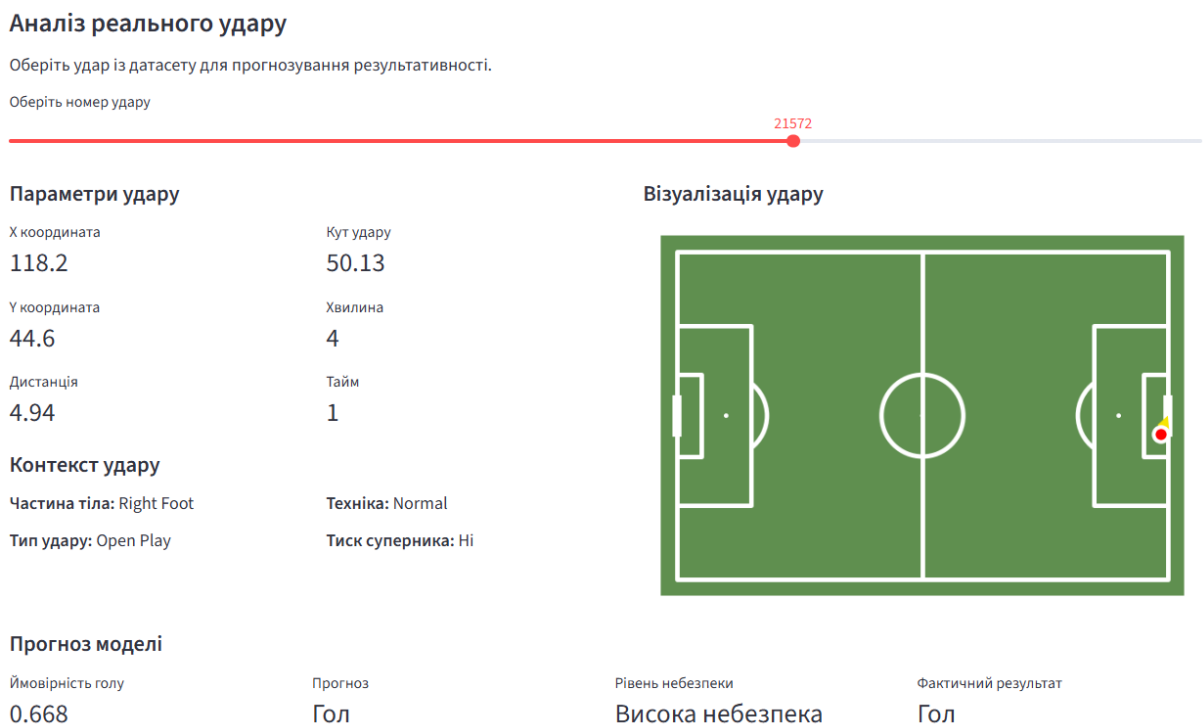


Рисунок 3.23 – Модуль аналізу реального удару

Для кожного обраного епізоду інтерфейс відображає основні просторові характеристики удару, включаючи координати положення гравця, дистанцію до воріт та кут удару. Додатково виводяться контекстні

параметри ігрової ситуації, зокрема тип удару, техніка виконання, частина тіла та наявність тиску з боку суперника. Такий підхід дозволяє користувачу аналізувати не лише фінальний прогноз моделі, а й фактори, що впливають на оцінювання небезпеки удару.

Після формування вхідного вектора ознак система передає дані до навченої моделі машинного навчання, яка виконує прогнозування ймовірності голу. Результат роботи моделі відображається у вигляді числової оцінки ймовірності, текстового прогнозу та рівня небезпеки ігрового епізоду.

Додатково у модулі реалізовано графічну візуалізацію удару на футбольному полі. Інтерфейс відображає положення удару та напрямок завершення атаки, що покращує інтерпретацію прогнозу та забезпечує наочне представлення ігрової ситуації.

Окремою особливістю модуля є можливість порівняння прогнозу моделі з фактичним результатом обраного удару з датасету. Це дозволяє оцінювати поведінку моделі на реальних ігрових ситуаціях та аналізувати коректність прогнозування.

Другим функціональним модулем системи є модуль моделювання ігрової ситуації, призначений для інтерактивного формування нових сценаріїв ударів та аналізу поведінки моделі прогнозування (рисунк 3.24). На відміну від модуля аналізу реальних ударів, у цьому режимі користувач самостійно задає параметри ігрової ситуації через елементи вебінтерфейсу.

У межах модуля реалізовано можливість введення просторових характеристик удару, включаючи координати положення гравця на полі, а також контекстних параметрів епізоду, зокрема типу удару, техніки виконання, частини тіла, хвилини матчу, тайму та наявності тиску з боку суперника. Для введення числових параметрів використовуються інтерактивні повзунки, а для категоріальних ознак – випадаючі списки.

Після введення параметрів система автоматично виконує розрахунок додаткових просторових ознак, зокрема дистанції до воріт та кута удару.

Сформований набір ознак передається до навченої моделі машинного навчання, після чого система автоматично виконує прогнозування ймовірності голу. Результат прогнозування відображається у вигляді числової оцінки ймовірності, текстового прогнозу та рівня небезпеки ігрової ситуації.

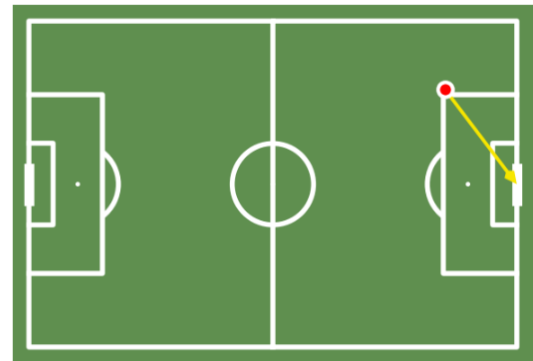
Модельовання ігрової ситуації

Введіть параметри удару для прогнозування ймовірності голу.

Параметри удару

X координата	102.36	Частина тіла	Head
Y координата	16.75	Тип удару	Corner
Хвилина	73	Техніка удару	Normal
Тайм	1	Тип атаки	Regular Play
Тиск суперника	0		

Візуалізація удару



Результат прогнозування

Ймовірність голу	Прогноз	Рівень небезпеки
0.438	Гол	Середня небезпека

Рисунок 3.24 – Модуль моделювання ігрової ситуації

Додатково у модулі реалізовано інтерактивну візуалізацію удару на футбольному полі. Положення гравця та напрямок завершення атаки автоматично оновлюються відповідно до введених користувачем координат, що забезпечує наочне представлення сформованої ігрової ситуації.

4 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

4.1 Результати навчання моделей машинного навчання

Після реалізації моделей машинного навчання було проведено експериментальне дослідження їх ефективності у задачі прогнозування результативності ударів. Метою даного етапу є оцінювання якості прогнозування, визначення сильних і слабких сторін окремих алгоритмів, а також вибір найбільш ефективного підходу для подальшого використання у розробленій системі.

4.1.1 Logistic Regression

Першою оцінювалася модель логістичної регресії, яка використовувалася як базовий алгоритм бінарної класифікації для прогнозування результативності ударів. Основні показники якості моделі наведено на рисунку 4.1.

	precision	recall	f1-score	support
Not Goal	0.96	0.82	0.88	5848
Goal	0.32	0.72	0.44	700
accuracy			0.81	6548
macro avg	0.64	0.77	0.66	6548
weighted avg	0.89	0.81	0.84	6548

Рисунок 4.1 – Classification report моделі Logistic Regression

З отриманих результатів видно, що модель забезпечує загальну точність класифікації на рівні 81%. Для класу нерезультативних ударів досягнуто високих значень precision (0.96), recall (0.82) та F1-score (0.88). Для класу результативних ударів значення precision становить 0.32, recall – 0.72, а F1-score – 0.44. Високе значення recall для класу Goal свідчить

про здатність моделі виявляти більшість результативних ударів, що є важливою характеристикою для задачі прогнозування голів за наявності дисбалансу класів.

Для більш детального аналізу помилок класифікації було побудовано матрицю помилок, наведену на рисунку 4.2.

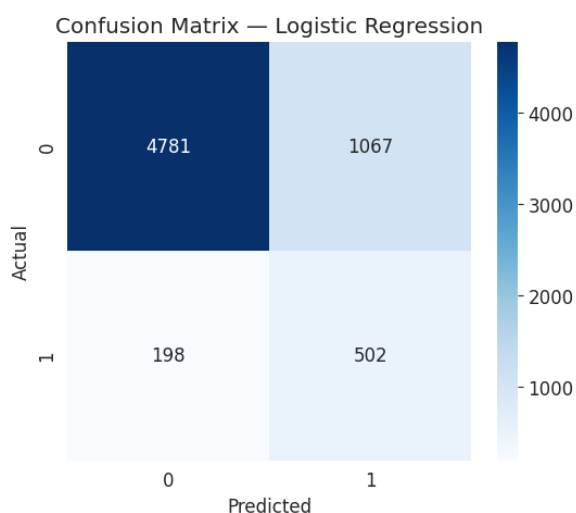


Рисунок 4.2 – Матриця помилок моделі Logistic Regression

Матриця помилок показує, що модель правильно класифікувала 4781 нерезультативний удар та 502 результативні удари. Водночас було отримано 1067 хибнопозитивних та 198 хибнонегативних прогнозів. Порівняно невелика кількість пропущених голів свідчить про орієнтацію моделі на виявлення результативних дій навіть ціною збільшення кількості хибних спрацьовувань.

Якість розділення класів додатково оцінювалася за допомогою ROC-кривої та Precision-Recall-кривої, наведених на рисунку 4.3.

Площа під ROC-кривою становить 0.836, що свідчить про добру здатність моделі відокремлювати результативні удари від нерезультативних. ROC-крива суттєво відхиляється від лінії випадкового класифікатора, підтверджуючи наявність закономірностей, виявлених під час навчання.

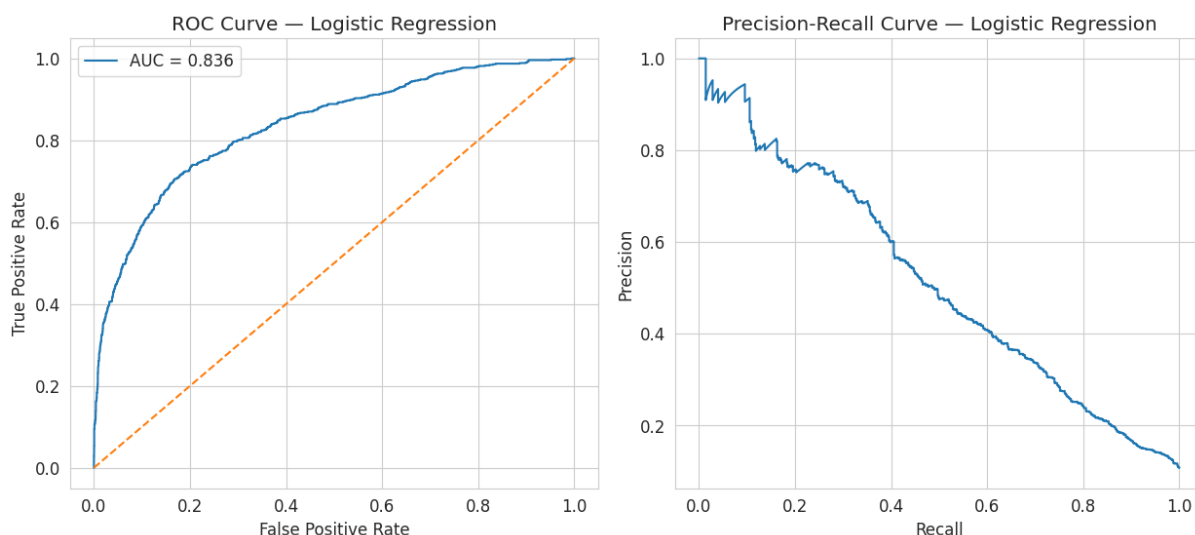


Рисунок 4.3 – ROC-крива та Precision-Recall-крива моделі Logistic Regression

Precision-Recall-крива дозволяє оцінити якість прогнозування саме результативних ударів, які є менш чисельним класом у вибірці. Отримана залежність демонструє поступове зниження точності зі зростанням повноти, що є типовою поведінкою для задач із дисбалансом класів.

Розподіл прогнозованих імовірностей для обох класів наведено на рисунку 4.4.

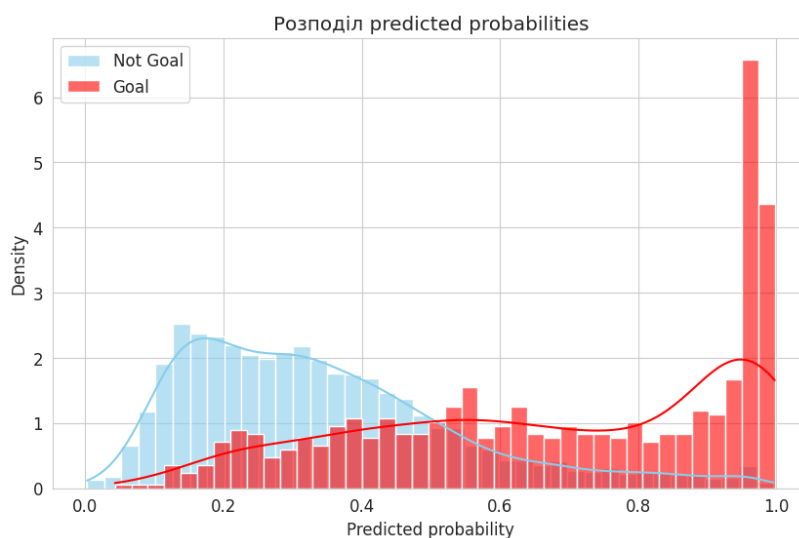


Рисунок 4.4 – Розподіл прогнозованих імовірностей моделі

З рисунка видно, що модель формує різні розподіли для результативних та нерезультативних ударів. Для більшості ударів класу Not Goal характерні низькі значення прогнозованої ймовірності, тоді як для класу Goal спостерігається зміщення розподілу у бік високих значень. Разом з тим між класами залишається область перекриття, яка і є основним джерелом помилок класифікації.

Для оцінювання узгодженості прогнозованих імовірностей із фактичною частотою настання події було побудовано графік калібрування, наведений на рисунку 4.5.

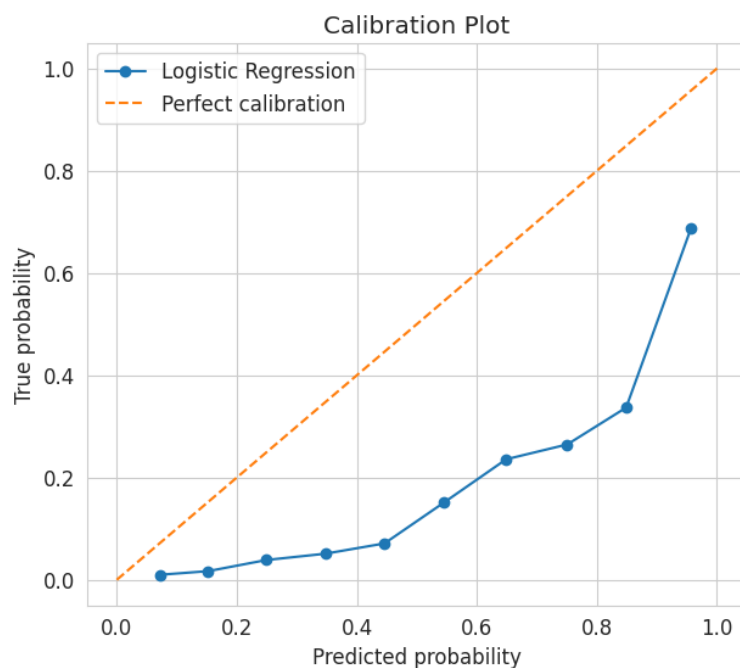


Рисунок 4.5 – Графік калібрування моделі Logistic Regression

Графік демонструє відхилення від лінії ідеального калібрування, особливо для високих значень прогнозованої ймовірності. Це свідчить про те, що модель не завжди коректно оцінює власну впевненість у прогнозах. Незважаючи на це, логістична регресія забезпечила стабільні результати та сформувала базовий рівень якості прогнозування, з яким надалі порівнювалися більш складні ансамблеві та нейромережеві моделі.

4.1.2 Random Forest

Наступним алгоритмом, який досліджувався у роботі, став Random Forest. Основні показники якості класифікації наведено на рисунку 4.6.

	precision	recall	f1-score	support
Not Goal	0.96	0.85	0.90	5848
Goal	0.35	0.68	0.46	700
accuracy			0.83	6548
macro avg	0.65	0.76	0.68	6548
weighted avg	0.89	0.83	0.85	6548

Рисунок 4.6 – Classification report моделі Random Forest

Модель забезпечила загальну точність класифікації на рівні 83%, що перевищує результат логістичної регресії. Для класу Goal отримано precision 0.35, recall 0.68 та F1-score 0.46, тоді як для класу Not Goal значення F1-score становить 0.90. Значення макро F1 дорівнює 0.68, що також перевищує показник базової моделі.

Матриця помилок моделі наведена на рисунку 4.7.

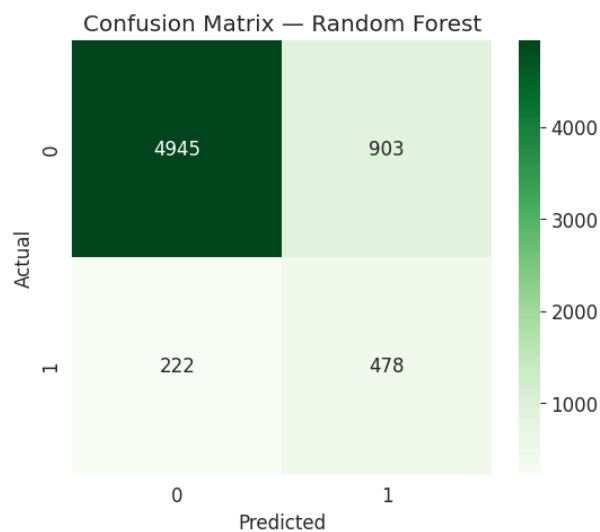


Рисунок 4.7 – Матриця помилок моделі Random Forest

Згідно з матрицею помилок, модель правильно класифікувала 4945 нерезультативних ударів та 478 результативних ударів. Кількість хибнопозитивних прогнозів зменшилася до 903 випадків порівняно з логістичною регресією, однак кількість пропущених голів дещо зростає і становить 222 випадки. Це свідчить про більш консервативну поведінку моделі під час прогнозування результативних ударів.

ROC-крива та Precision-Recall-крива моделі наведені на рисунку 4.8.

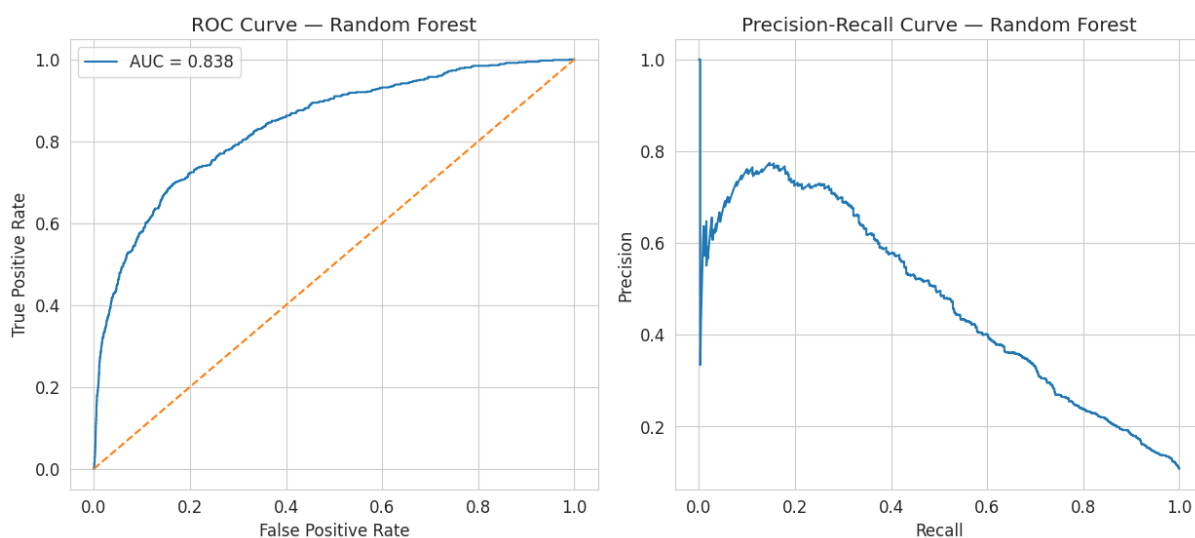


Рисунок 4.8 – ROC-крива та Precision-Recall-крива моделі Random Forest

Площа під ROC-кривою становить 0.838, що є дещо вищим результатом порівняно з логістичною регресією. Крива демонструє стабільне розділення класів на всьому діапазоні порогів класифікації.

Precision-Recall-крива також характеризується кращими показниками точності у зоні середніх значень recall. Це свідчить про здатність ансамблевої моделі ефективніше відокремлювати результативні удари від нерезультативних у складних випадках, де лінійна модель демонструє нижчу впевненість.

Розподіл прогнозованих імовірностей наведено на рисунку 4.9.

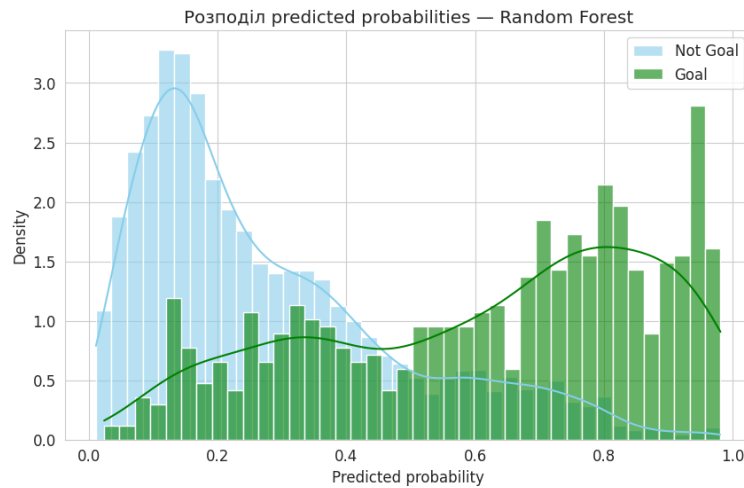


Рисунок 4.9 – Розподіл прогнозованих імовірностей моделі Random Forest

На відміну від логістичної регресії, модель Random Forest формує більш виражене розділення між класами. Для нерезультативних ударів основна частина прогнозів концентрується в області низьких значень імовірності, тоді як для результативних ударів спостерігається зміщення до діапазону високих імовірностей. Менша область перекриття між розподілами пояснює покращення показників класифікації.

Графік калібрування моделі наведено на рисунку 4.10.

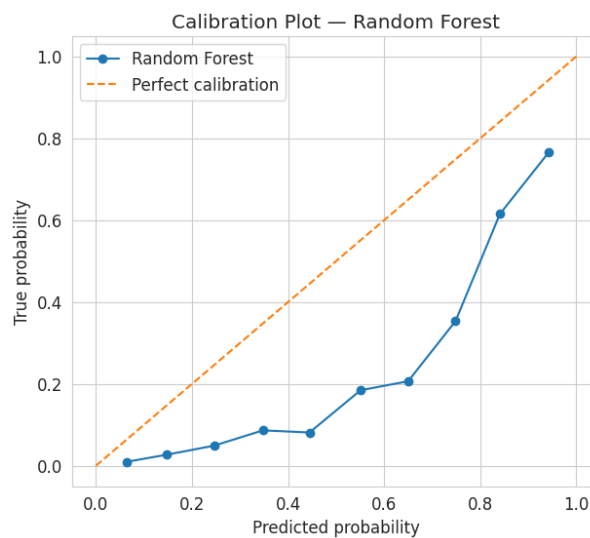


Рисунок 4.10 – Графік калібрування моделі Random Forest

Отримана залежність демонструє певне покращення калібрування порівняно з логістичною регресією, особливо для середніх значень прогнозованої ймовірності. Водночас для високих значень ймовірності все ще спостерігається відхилення від лінії ідеального калібрування, що вказує на схильність моделі до переоцінювання окремих прогнозів. Загалом Random Forest продемонструвала більш збалансовані результати класифікації та покращення більшості ключових метрик порівняно з базовою лінійною моделлю.

4.1.3 XGBoost

Для подальшого підвищення якості прогнозування було використано модель XGBoost, яка належить до ансамблевих методів градієнтного бустингу. Під час налаштування моделі додатково виконувався підбір порогу класифікації, що дозволило адаптувати прогнозування до умов дисбалансу класів. Основні показники якості моделі наведено на рисунку 4.11.

	precision	recall	f1-score	support
Not Goal	0.94	0.92	0.93	5848
Goal	0.45	0.52	0.48	700
accuracy			0.88	6548
macro avg	0.70	0.72	0.71	6548
weighted avg	0.89	0.88	0.88	6548

Рисунок 4.11 – Classification report моделі XGBoost

Отримані результати демонструють найвищу серед розглянутих моделей загальну точність класифікації – 88%. Для класу Goal значення precision становить 0.45, recall – 0.52, а F1-score – 0.48. Значення macro F1 досягло 0.71, що перевищує результати як логістичної регресії, так і Random Forest. Це свідчить про більш збалансоване прогнозування обох класів.

Матриця помилок моделі наведена на рисунку 4.12.

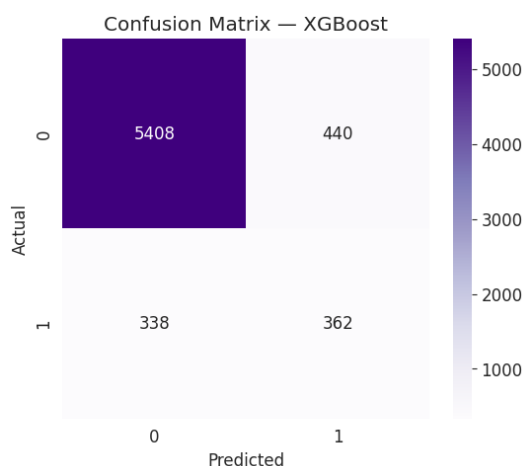


Рисунок 4.12 – Матриця помилок моделі XGBoost

З матриці помилок видно, що модель правильно класифікувала 5408 нерезультативних ударів та 362 результативні удари. Кількість хибнопозитивних прогнозів скоротилася до 440 випадків, що майже вдвічі менше порівняно з Random Forest. При цьому кількість пропущених голів становить 338 випадків. Така поведінка свідчить про більш обережне віднесення ударів до класу Goal та зменшення кількості помилкових спрацьовувань.

ROC-крива та Precision-Recall-крива моделі наведені на рисунку 4.13.

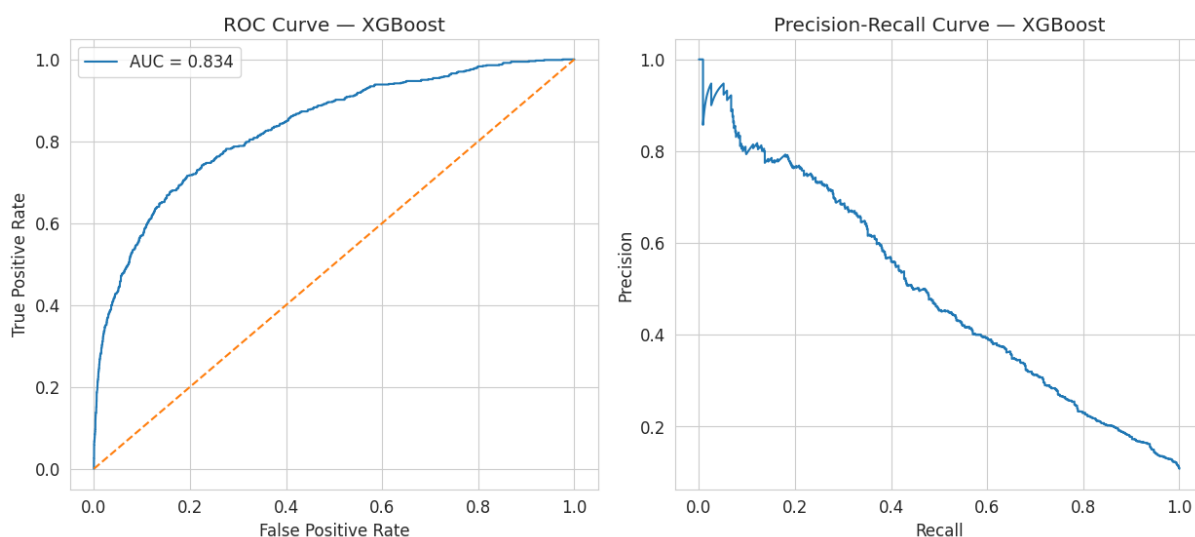


Рисунок 4.13 – ROC-крива та Precision-Recall-крива моделі XGBoost

Площа під ROC-кривою становить 0.834. Незважаючи на те, що це значення близьке до результатів попередніх моделей, Precision-Recall-крива демонструє більш високі значення precision на значній частині діапазону recall. Це свідчить про кращу якість ранжування потенційно результативних ударів та більш впевнене розділення класів.

Розподіл прогнозованих імовірностей наведено на рисунку 4.14.

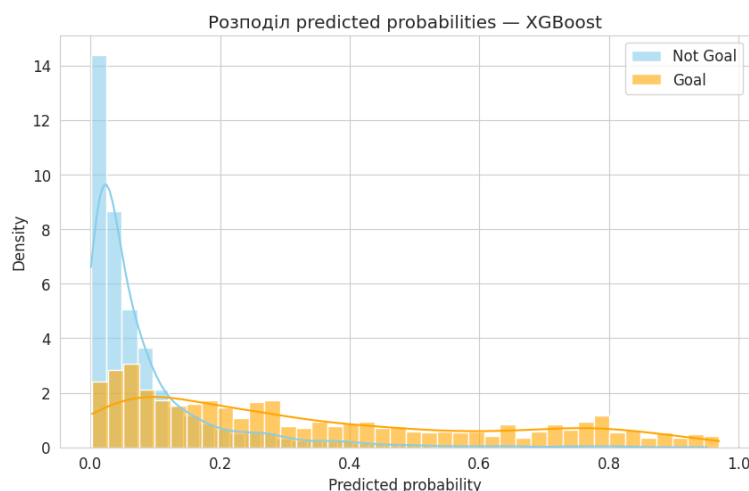


Рисунок 4.14 – Розподіл прогнозованих імовірностей моделі XGBoost

З рисунка видно, що модель формує найбільш виражене розділення між класами серед розглянутих алгоритмів. Для нерезультативних ударів переважна більшість прогнозів концентрується поблизу нульових значень імовірності, тоді як для результативних ударів спостерігається широкий розподіл із суттєвою концентрацією у зоні високих значень. Це свідчить про високу здатність моделі відокремлювати перспективні гольові моменти від звичайних ударів.

Графік калібрування моделі наведено на рисунку 4.15. На відміну від попередніх моделей, крива калібрування XGBoost проходить поблизу лінії ідеального калібрування практично на всьому діапазоні прогнозованих імовірностей. Це свідчить про високу узгодженість між прогнозованими значеннями та фактичною частотою настання голів. Отримані результати

підтверджують, що XGBoost є одним із найбільш ефективних підходів для задачі прогнозування результативності ударів та забезпечує високий рівень якості як класифікації, так і оцінювання ймовірностей.

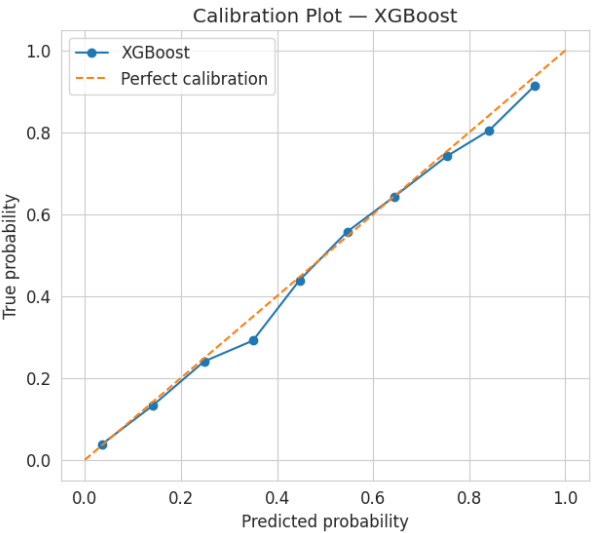


Рисунок 4.15 – Графік калібрування моделі XGBoost

4.1.4 CatBoost

Ще одним алгоритмом градієнтного бустингу, дослідженим у роботі, стала модель CatBoost. Аналогічно до XGBoost, для неї виконувався підбір оптимального порогу класифікації з урахуванням дисбалансу класів. Основні показники якості моделі наведено на рисунку 4.16.

	precision	recall	f1-score	support
Not Goal	0.94	0.93	0.94	5848
Goal	0.47	0.51	0.49	700
accuracy			0.89	6548
macro avg	0.71	0.72	0.71	6548
weighted avg	0.89	0.89	0.89	6548

Рисунок 4.16 – Classification report моделі CatBoost

Отримані результати демонструють найвищу серед усіх розглянутих моделей загальну точність класифікації – 89%. Для класу Goal значення precision становить 0.47, recall – 0.51, а F1-score – 0.49. Значення macro F1 дорівнює 0.71, що відповідає результату XGBoost. Водночас CatBoost забезпечує дещо вищу точність прогнозування результативних ударів, що свідчить про краще відокремлення потенційно гольових моментів від нерезультативних.

Матриця помилок моделі наведена на рисунку 4.17.

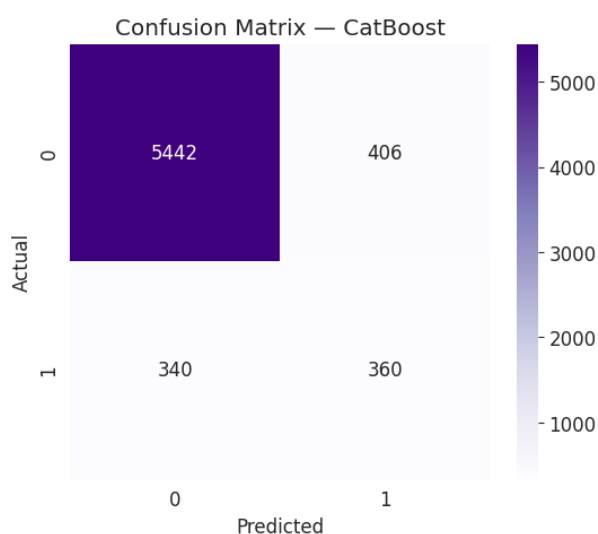


Рисунок 4.17 – Матриця помилок моделі CatBoost

З матриці помилок видно, що модель правильно класифікувала 5442 нерезультативні удари та 360 результативних ударів. Кількість хибнопозитивних прогнозів скоротилася до 406 випадків, що є найкращим результатом серед розглянутих моделей. При цьому кількість пропущених голів становить 340 випадків і практично не відрізняється від результату XGBoost. Таким чином, CatBoost досягла найбільш збалансованого співвідношення між кількістю хибних спрацьовувань та пропущених результативних ударів.

ROC-крива та Precision-Recall-крива моделі наведені на рисунку 4.18.

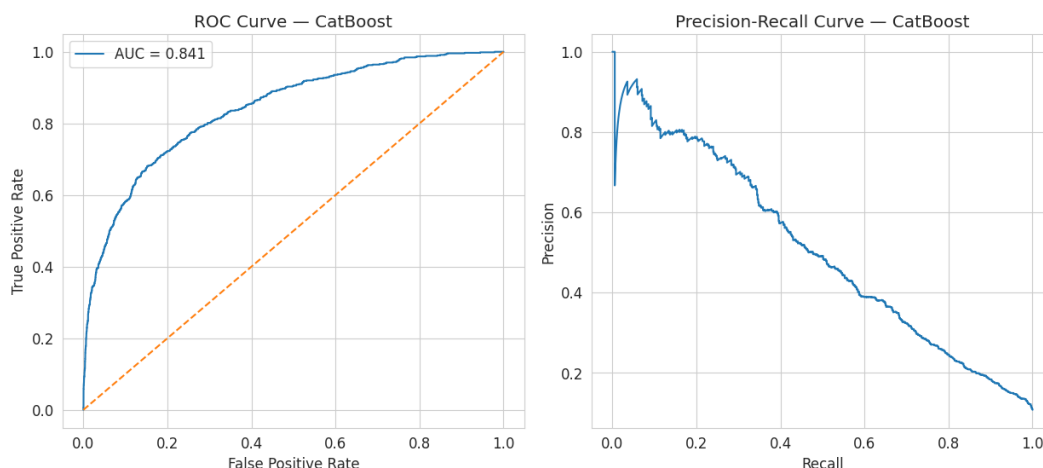


Рисунок 4.18 – ROC-крива та Precision-Recall-крива моделі CatBoost

Площа під ROC-кривою становить 0.841, що є найвищим показником серед усіх досліджених моделей. ROC-крива демонструє стабільно високу здатність відокремлювати результативні удари від нерезультативних на різних значеннях порогу класифікації.

Precision-Recall-крива також перевищує результати попередніх моделей на значній частині діапазону recall, що підтверджує ефективність алгоритму при роботі з незбалансованими даними та складними залежностями між ознаками.

Розподіл прогнозованих імовірностей наведено на рисунку 4.19.

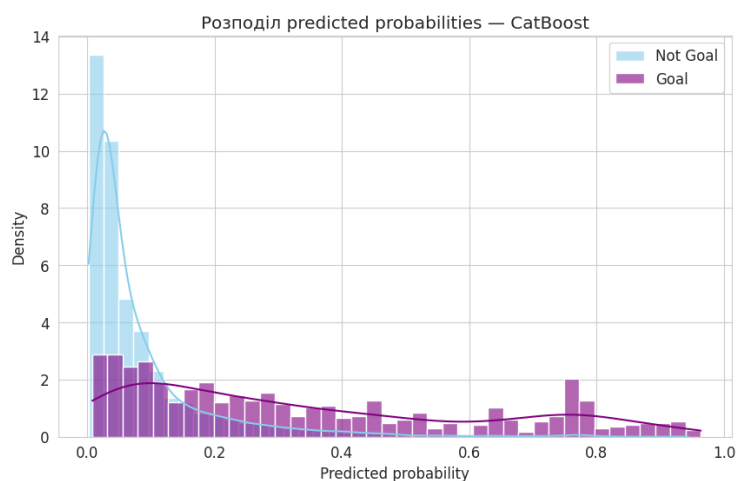


Рисунок 4.19 – Розподіл прогнозованих імовірностей моделі CatBoost

Як і у випадку XGBoost, модель формує чітке розділення між класами. Основна маса нерезультативних ударів отримує низькі значення прогнозованої ймовірності, тоді як результативні удари зміщені у бік середніх та високих значень. Порівняно невелика область перекриття між розподілами пояснює високі показники точності та F1-міри.

Графік калібрування моделі наведено на рисунку 4.20.

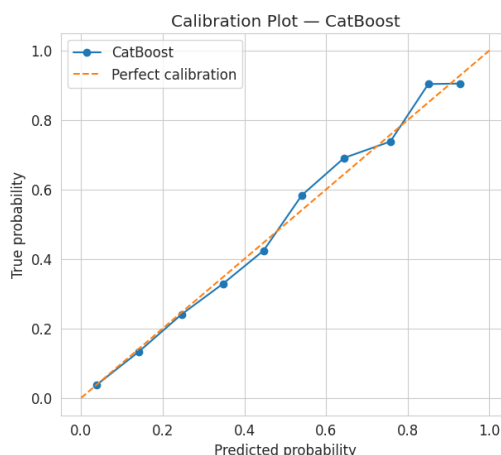


Рисунок 4.20 – Графік калібрування моделі CatBoost

Крива калібрування проходить дуже близько до лінії ідеального калібрування практично на всьому діапазоні прогнозованих імовірностей. Це свідчить про високу узгодженість між прогнозованими значеннями та реальною частотою настання події.

4.1.5 MLP

Для оцінювання можливостей нейромережевого підходу у задачі прогнозування результативності ударів було використано багат шаровий перцептрон (MLP). На відміну від попередніх моделей, MLP виконує нелінійне моделювання залежностей між ознаками та здатний враховувати складні взаємозв'язки між характеристиками удару. Основні показники якості моделі наведено на рисунку 4.21.

	precision	recall	f1-score	support
Not Goal	0.92	0.96	0.94	5848
Goal	0.51	0.32	0.40	700
accuracy			0.90	6548
macro avg	0.72	0.64	0.67	6548
weighted avg	0.88	0.90	0.88	6548

Рисунок 4.21 – Classification report моделі MLP

Модель продемонструвала найвищу загальну точність класифікації серед усіх досліджених алгоритмів – 90%. Для класу Goal значення precision становить 0.51, recall – 0.32, а F1-score – 0.40. Незважаючи на високе значення precision, модель виявилася менш ефективною у виявленні результативних ударів, про що свідчить суттєве зниження recall порівняно з ансамблевими моделями.

Матриця помилок моделі наведена на рисунку 4.22.

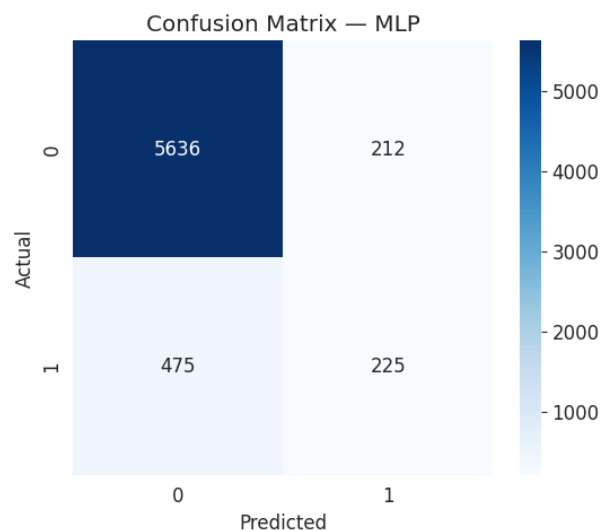


Рисунок 4.22 – Матриця помилок моделі MLP

З матриці помилок видно, що модель правильно класифікувала 5636 нерезультативних ударів та лише 225 результативних ударів. Кількість хибнопозитивних прогнозів становить 212 випадків, що є найменшим значенням серед усіх досліджених моделей. Водночас кількість пропущених

голів зростає до 475 випадків. Таким чином, MLP демонструє дуже обережну стратегію прогнозування, відносячи до класу Goal лише найбільш очевидні голюві ситуації.

ROC-крива та Precision-Recall-крива моделі наведені на рисунку 4.23.

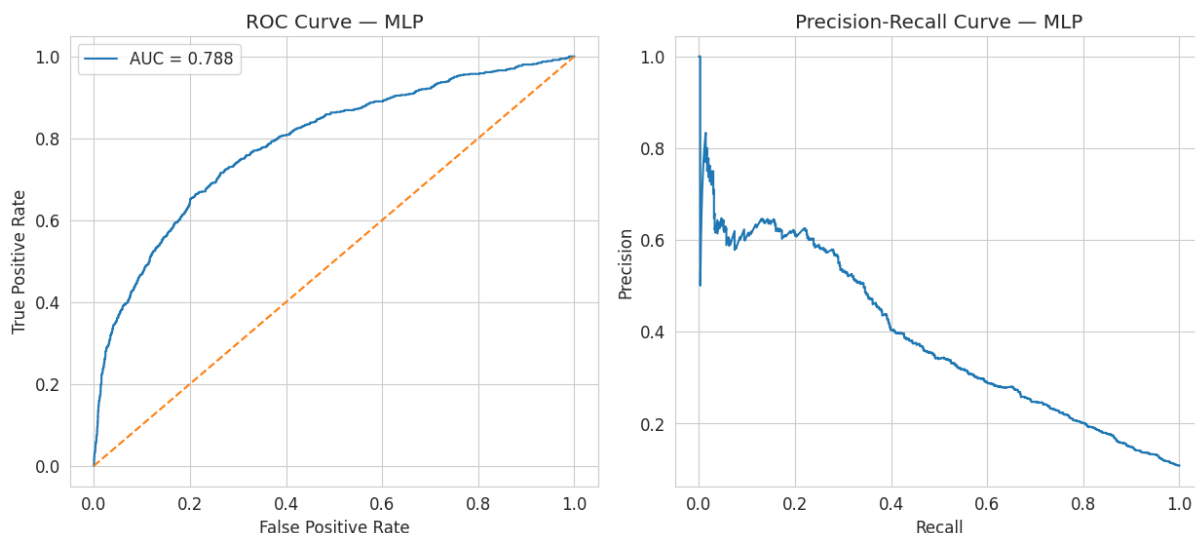


Рисунок 4.23 – ROC-крива та Precision-Recall-крива моделі MLP

Площа під ROC-кривою становить 0.788, що є найнижчим результатом серед розглянутих моделей. Аналогічна тенденція спостерігається і для Precision-Recall-кривої, яка проходить нижче кривих ансамблевих алгоритмів на більшій частині діапазону recall. Це свідчить про обмежену здатність моделі відокремлювати результативні удари від нерезультативних у межах сформованого набору ознак.

Розподіл прогнозованих імовірностей наведено на рисунку 4.24.

З рисунка видно, що модель концентрує значну частину прогнозів поблизу нульових значень імовірності. Для результативних ударів спостерігається зміщення розподілу у бік більших значень, однак область перекриття між класами залишається суттєвою. Це пояснює низьку повноту прогнозування голів та значну кількість пропущених результативних ударів.

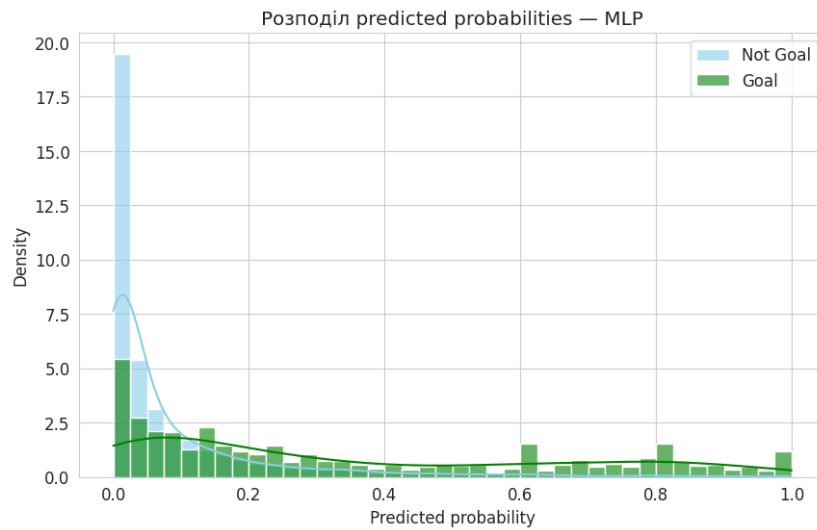


Рисунок 4.24 – Розподіл прогнозованих імовірностей моделі MLP

Графік калібрування моделі наведено на рисунку 4.25.

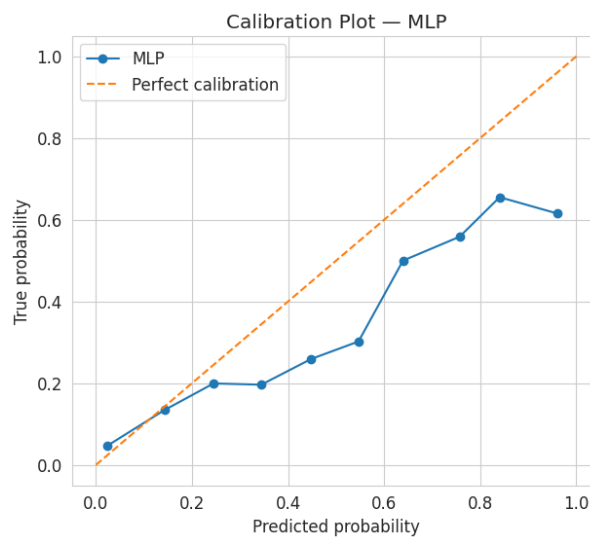


Рисунок 4.25 – Графік калібрування моделі MLP

Крива калібрування демонструє помітні відхилення від лінії ідеального калібрування, особливо для високих значень прогнозованої ймовірності. Це свідчить про те, що оцінки ймовірностей, сформовані нейромережею, не завжди відповідають фактичній частоті настання події.

4.2 Порівняльний аналіз моделей прогнозування

Після окремого оцінювання кожної моделі було виконано узагальнене порівняння отриманих результатів. Для цього використовувалися основні метрики класифікації, показники якості ймовірного прогнозування, а також графіки, що дозволяють порівняти здатність моделей розділяти класи та формувати узгоджені оцінки ймовірності голу.

Порівняльну таблицю метрик моделей наведено на рисунку 4.26.

	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Log Loss	Brier Score
0	Logistic Regression	0.807	0.320	0.717	0.442	0.836	0.499	0.158
1	Random Forest	0.828	0.346	0.683	0.459	0.838	0.415	0.129
2	XGBoost	0.881	0.451	0.517	0.482	0.834	0.253	0.072
3	CatBoost	0.886	0.470	0.514	0.491	0.841	0.249	0.071
4	MLP	0.895	0.515	0.321	0.396	0.788	0.316	0.084

Рисунок 4.26 – Порівняльна таблиця метрик моделей прогнозування

Загальна точність моделей знаходиться в діапазоні від 0.807 до 0.895. Найвище значення асигасу отримано для MLP, однак цей показник не є достатнім для вибору найкращої моделі через дисбаланс класів. Для задачі прогнозування голів важливими є не лише правильне визначення нерезультативних ударів, а й здатність моделі виявляти клас Goal.

За F1-score для класу Goal найкращий результат показала модель CatBoost – 0.491. Близький результат отримано для XGBoost – 0.482. Logistic Regression та Random Forest мають нижчі значення F1-score, а MLP, попри найвищу асигасу, демонструє найнижчий баланс між precision і recall для результативних ударів. Це свідчить про те, що нейромережева модель частіше уникає прогнозу класу Goal, зменшуючи кількість хибнопозитивних спрацьовувань, але пропускаючи значну частину голів.

За ROC-AUC найкращий результат отримано для CatBoost – 0.841. Показники Logistic Regression, Random Forest та XGBoost є близькими, тоді як MLP має помітно нижче значення ROC-AUC. Додатково CatBoost і XGBoost показали найнижчі значення LogLoss та Brier Score, що свідчить про кращу якість імовірнісних прогнозів.

Порівняння ROC-кривих моделей наведено на рисунку 4.27.

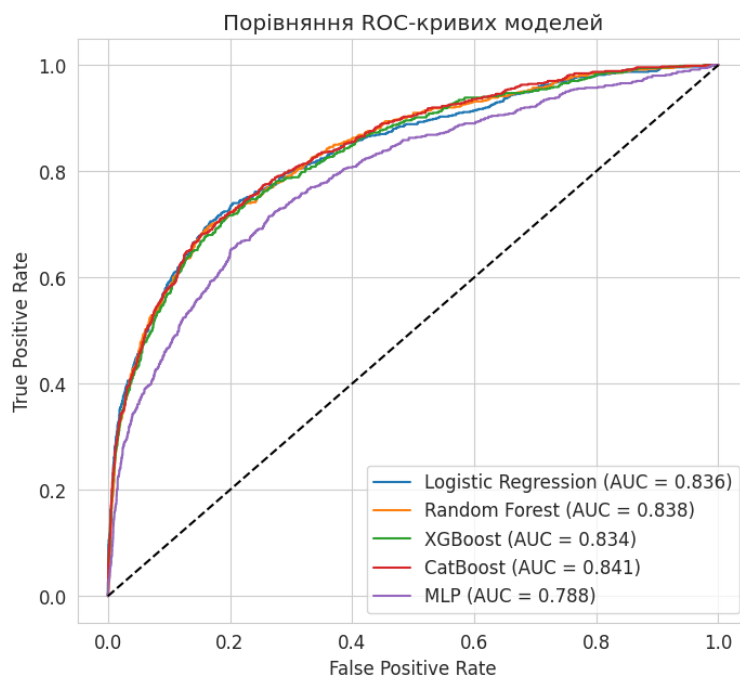


Рисунок 4.27 – Порівняння ROC-кривих моделей прогнозування

ROC-криві Logistic Regression, Random Forest, XGBoost та CatBoost розташовані близько одна до одної, що підтверджує подібний рівень здатності цих моделей до загального розділення класів. Водночас CatBoost має найвище значення AUC, а MLP помітно поступається іншим моделям майже на всьому діапазоні значень False Positive Rate.

Порівняння Precision-Recall-кривих наведено на рисунку 4.28.

Precision-Recall-криві краще відображають поведінку моделей для менш чисельного класу Goal. За цими кривими видно, що CatBoost, XGBoost, Random Forest та Logistic Regression мають близьку поведінку в

середній частині діапазону recall, тоді як MLP демонструє нижчі значення precision при зростанні повноти. Це узгоджується з результатами classification report, де MLP має найменше значення recall для класу Goal.

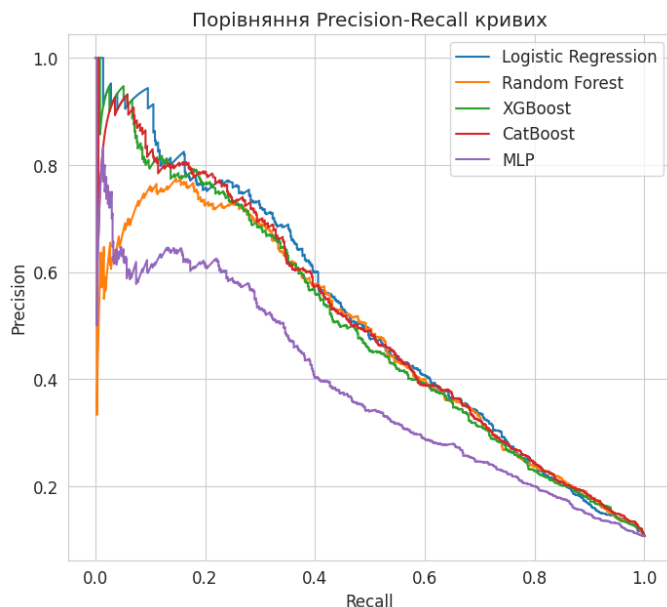


Рисунок 4.28 – Порівняння Precision-Recall-кривих моделей прогнозування

Порівняння калібрування моделей наведено на рисунку 4.29.

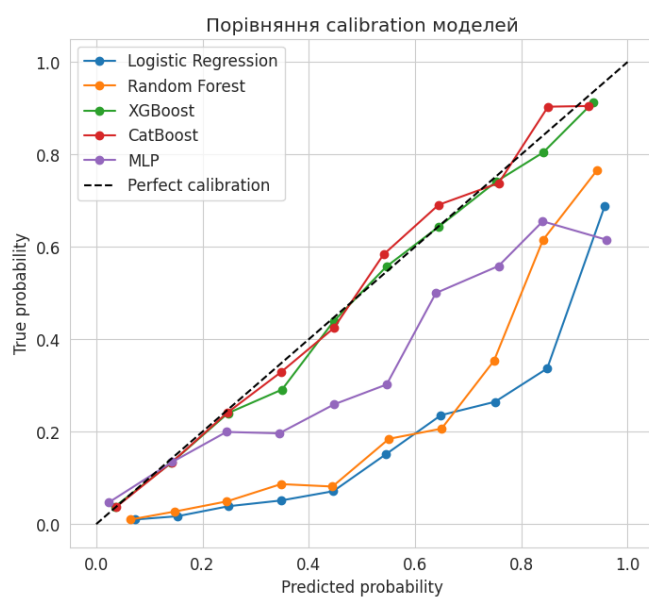


Рисунок 4.29 – Порівняння калібрування моделей прогнозування

Аналіз калібрування показав, що найближче до лінії ідеального калібрування проходять криві CatBoost та XGBoost. Це означає, що прогнозовані ними ймовірності найкраще узгоджуються з фактичною частотою голів. Logistic Regression та Random Forest мають помітні відхилення, особливо для середніх і високих значень прогнозованої ймовірності. MLP також демонструє нестабільне калібрування, що обмежує надійність її ймовірнісних оцінок.

Проведене порівняння показало, що найкращі результати серед досліджуваних моделей продемонструвала модель CatBoost, яка забезпечила найвищі значення F1-міри, ROC-AUC та найменші значення Log Loss і Brier Score. Високі результати також показали моделі XGBoost та Random Forest, тоді як Logistic Regression і MLP поступилися ансамблевим методам за більшістю показників якості.

4.3 Аналіз роботи системи

Після оцінювання та порівняння моделей прогнозування було проведено аналіз роботи розробленої системи на практичних прикладах. Метою даного етапу є перевірка того, наскільки отримані моделі здатні коректно оцінювати як реальні ігрові епізоди з датасету, так і штучно сформовані користувачем ситуації.

4.3.1 Аналіз прогнозування на реальних ігрових ситуаціях

Для оцінювання роботи розробленої системи було проаналізовано прогнозування результативності на реальних ігрових епізодах із тестової вибірки. Розглянуто як приклади коректних прогнозів, так і ситуації, у яких модель припустилася помилки. Такий підхід дозволяє оцінити поведінку системи в умовах, наближених до практичного використання.

Приклад коректного прогнозування результативного удару наведено на рисунку 4.30. У даному випадку удар виконувався з близької дистанції до воріт (3.5 м), а модель оцінила ймовірність голу на рівні 0.926. Прогноз системи співпав із фактичним результатом епізоду, що свідчить про коректне визначення високонебезпечної голювальної ситуації.

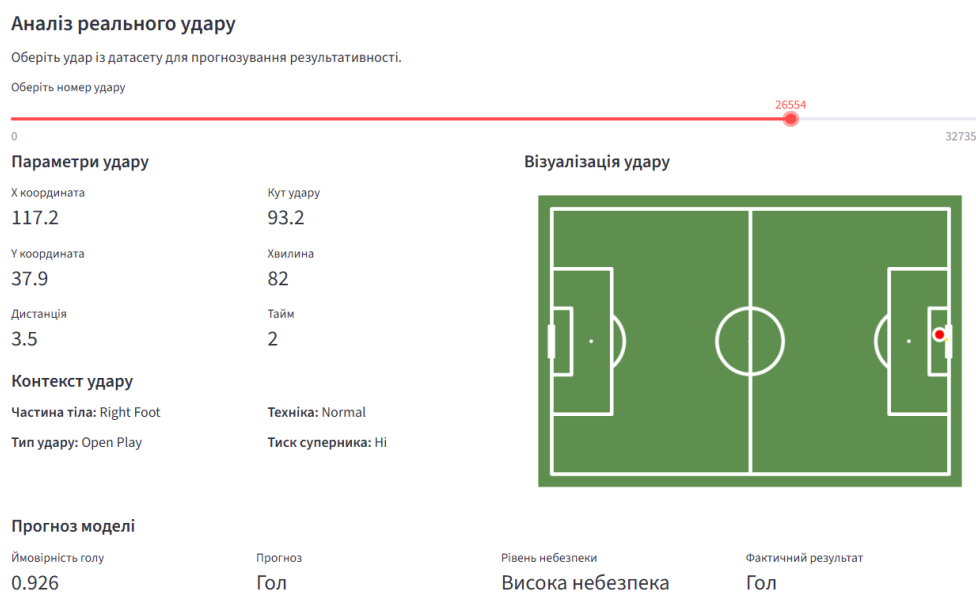


Рисунок 4.30 – Приклад коректного прогнозування результативного удару

На рисунку 4.31 представлено приклад коректного прогнозування нерезультативного удару. Удар виконувався з відносно великої дистанції (22.72 м) та під малим кутом до воріт, унаслідок чого модель визначила низьку ймовірність голу (0.026) та віднесла епізод до класу «Не гол». Фактичний результат повністю підтвердив прогноз системи.

Приклад хибнопозитивного прогнозу наведено на рисунку 4.32. Фактично удар не завершився голом, однак модель оцінила ймовірність результативного завершення на рівні 0.466 і віднесла епізод до класу «Гол». Такий прогноз не є випадковою помилкою: удар виконувався з близької дистанції 7.64 м, без тиску суперника, з відкритої гри та з достатньо вигідного кута 41.34°. За наявними ознаками епізод справді має характеристики небезпечного моменту. Помилка виникає через те, що

модель не враховує фактори, які могли визначити фактичний результат: точність удару, положення воротаря, розташування захисників, швидкість м'яча або якість завершення атаки.

Аналіз реального удару

Оберіть удар із датасету для прогнозування результативності.

Оберіть номер удару

24259

Параметри удару

X координата

101.3

Y координата

27.1

Дистанція

22.72

Контекст удару

Частина тіла: Left Foot

Тип удару: Open Play

Кут удару

15.23

Хвилина

91

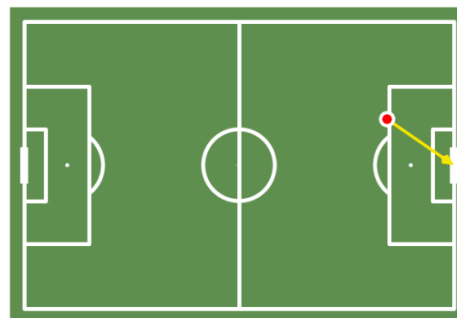
Тайм

2

Техніка: Normal

Тиск суперника: Ні

Візуалізація удару



Прогноз моделі

Ймовірність голу

0.026

Прогноз

Не гол

Рівень небезпеки

Низька небезпека

Фактичний результат

Не гол

Рисунок 4.31 – Приклад коректного прогнозування нерезультативного удару

Аналіз реального удару

Оберіть удар із датасету для прогнозування результативності.

Оберіть номер удару

26387

0

32735

Параметри удару

X координата

114.6

Y координата

34.6

Дистанція

7.64

Контекст удару

Частина тіла: Left Foot

Тип удару: Open Play

Кут удару

41.34

Хвилина

52

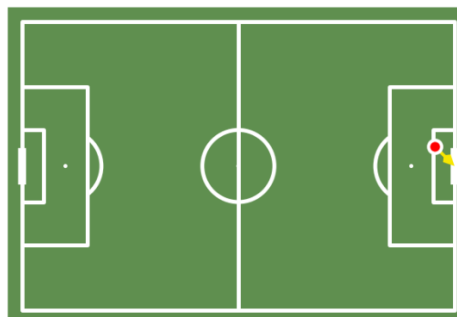
Тайм

2

Техніка: Normal

Тиск суперника: Ні

Візуалізація удару



Прогноз моделі

Ймовірність голу

0.466

Прогноз

Гол

Рівень небезпеки

Середня небезпека

Фактичний результат

Не гол

Рисунок 4.32 – Приклад хибнопозитивного прогнозу системи

На рисунку 4.33 показано приклад хибнонегативного прогнозу. Модель оцінила ймовірність голу лише у 0.084 і віднесла удар до класу «Не гол», однак фактично атака завершилася взяттям воріт. З позиції доступних моделі ознак такий прогноз є зрозумілим: удар виконувався з дистанції 19.1 м, під тиском суперника, а кут удару становив 19.39° , тобто позиція не належала до очевидно високонебезпечних. Фактичний гол у такій ситуації міг бути зумовлений індивідуальною майстерністю виконавця, помилкою воротаря, траєкторією удару або іншими динамічними факторами, які не представлені у сформованому датасеті.

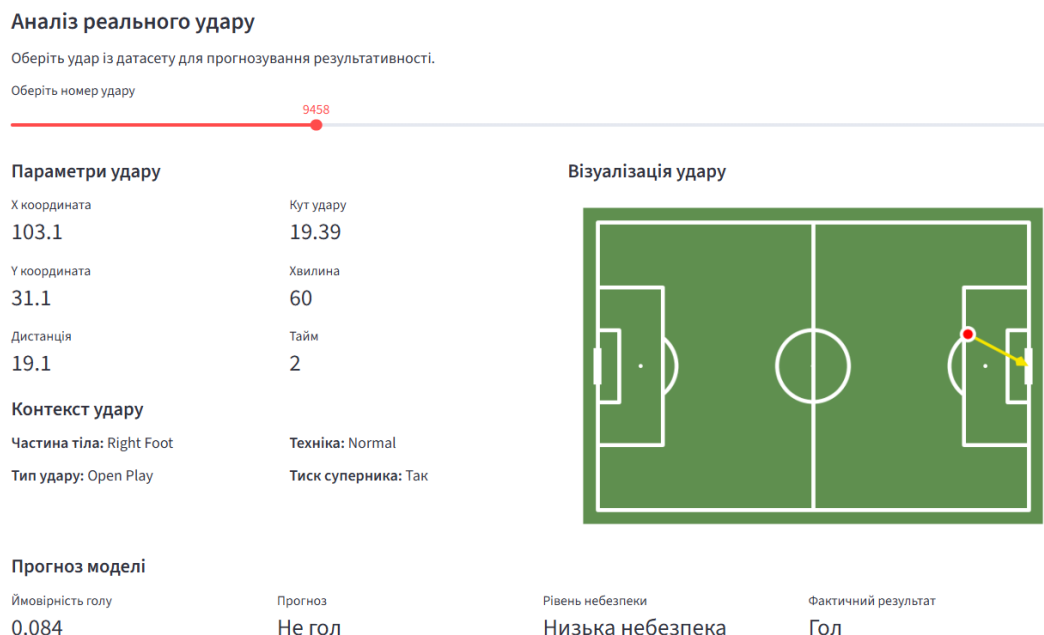


Рисунок 4.33 – Приклад хибнонегативного прогнозу системи

Проведений аналіз показав, що система успішно ідентифікує більшість типових результативних та нерезультативних ударів. Водночас окремі помилки виникають у прикордонних ситуаціях, де набір ознак не дозволяє однозначно відокремити голевий епізод від неголевого. Це відповідає результатам кількісного оцінювання моделей, отриманим у попередніх підрозділах.

4.3.2 Аналіз роботи системи при моделюванні ударів

Для оцінювання впливу дистанції до воріт на прогноз системи було змодельовано серію ударів із однаковими характеристиками. Незмінними залишалися координата Y , хвилина матчу, тайм, тип удару, техніка виконання, частина тіла, тип атаки та наявність тиску суперника. Змінювалася лише координата X , що безпосередньо впливало на відстань до воріт.

Результати моделювання наведено на рисунку 4.34. У першому випадку удар виконувався з близької дистанції до воріт, унаслідок чого система оцінила ймовірність голу на рівні 0.494 та віднесла епізод до категорії середньої небезпеки. Після збільшення дистанції прогнозована ймовірність результативного завершення атаки знизилася до 0.261. Подальше віддалення точки удару від воріт призвело до зменшення ймовірності голу до 0.089 та зміни прогнозу на клас «Не гол».

Отримані результати демонструють, що модель адекватно реагує на зміну просторових характеристик удару. Зі збільшенням дистанції до воріт прогнозована ймовірність голу поступово зменшується, що відповідає закономірностям, які спостерігаються у реальних футбольних матчах та використовуються в сучасних моделях оцінювання результативності ударів.

У другому сценарії було змодельовано зміну положення удару відносно центральної осі воріт за незмінної координати X та інших параметрів епізоду. Результати наведено на рисунку 4.35.

Порівняння двох змодельованих ситуацій показало, що прогнозована ймовірність голу не змінюється лінійно лише залежно від кута удару. У першому випадку, коли удар виконувався з більш бокового положення, система оцінила ймовірність голу на рівні 0.732. У другому випадку, при розташуванні ближче до центральної зони, ймовірність становила 0.422. Це свідчить про те, що модель враховує сукупний вплив просторових і контекстних характеристик удару, а не окремі ознаки ізольовано.

Моделювання ігрової ситуації

Введіть параметри удару для прогнозування ймовірності голу.

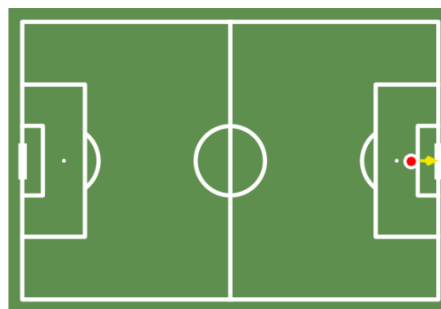
Параметри удару

X координата	112.00	Частина тіла	Right Foot
Y координата	40.00	Тип удару	Open Play
Хвилина	60	Техніка удару	Normal
Тайм	2	Тип атаки	Regular Play
Тиск суперника	0		

Результат прогнозування

Ймовірність голу	Прогноз	Рівень безпеки
0.494	Гол	Середня небезпека

Візуалізація удару



Моделювання ігрової ситуації

Введіть параметри удару для прогнозування ймовірності голу.

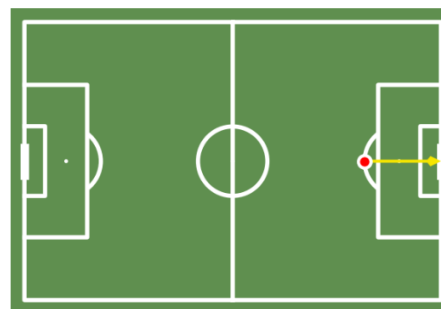
Параметри удару

X координата	97.85	Частина тіла	Right Foot
Y координата	40.00	Тип удару	Open Play
Хвилина	60	Техніка удару	Normal
Тайм	2	Тип атаки	Regular Play
Тиск суперника	0		

Результат прогнозування

Ймовірність голу	Прогноз	Рівень безпеки
0.261	Гол	Середня небезпека

Візуалізація удару



Моделювання ігрової ситуації

Введіть параметри удару для прогнозування ймовірності голу.

Параметри удару

X координата	89.73	Частина тіла	Right Foot
Y координата	40.00	Тип удару	Open Play
Хвилина	60	Техніка удару	Normal
Тайм	2	Тип атаки	Regular Play
Тиск суперника	0		

Результат прогнозування

Ймовірність голу	Прогноз	Рівень безпеки
0.089	Не гол	Низька небезпека

Візуалізація удару

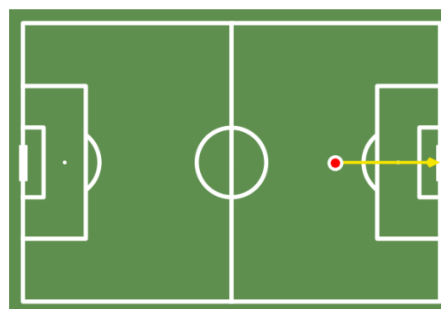


Рисунок 4.34 – Вплив збільшення дистанції до воріт на прогнозовану ймовірність голу

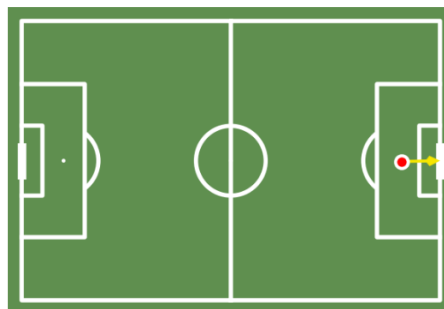
Моделювання ігрової ситуації

Введіть параметри удару для прогнозування ймовірності голу.

Параметри удару

X координата	109.00	Частина тіла	Right Foot
Y координата	40.22	Тип удару	Open Play
Хвилина	60	Техніка удару	Normal
Тайм	2	Тип атаки	Regular Play
Тиск суперника	0		

Візуалізація удару



Результат прогнозування

Ймовірність голу	Прогноз	Рівень небезпеки
0.422	Гол	Середня небезпека

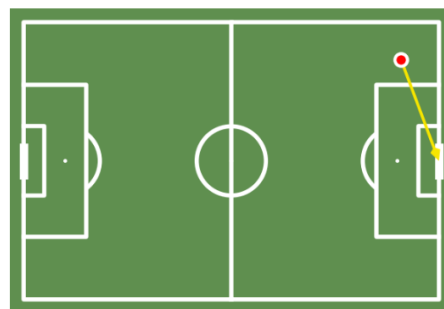
Моделювання ігрової ситуації

Введіть параметри удару для прогнозування ймовірності голу.

Параметри удару

X координата	109.00	Частина тіла	Right Foot
Y координата	10.63	Тип удару	Open Play
Хвилина	60	Техніка удару	Normal
Тайм	2	Тип атаки	Regular Play
Тиск суперника	0		

Візуалізація удару



Результат прогнозування

Ймовірність голу	Прогноз	Рівень небезпеки
0.732	Гол	Висока небезпека

Рисунок 4.35 – Вплив зміни положення удару відносно центральної осі воріт на прогноз системи

Результати моделювання демонструють, що система враховує взаємозв'язок між просторовими характеристиками удару та ймовірністю взяття воріт. Найбільший вплив на прогноз мають параметри, безпосередньо пов'язані з розташуванням гравця відносно воріт, що відповідає висновкам, отриманим під час аналізу моделей машинного навчання. Це підтверджує практичну придатність розробленої системи для аналізу та оцінювання потенційної результативності футбольних ударів.

ВИСНОВКИ

У кваліфікаційній роботі розв'язано актуальну задачу розробки інтелектуальної системи прогнозування результативних дій у спортивній аналітиці на основі методів машинного навчання. Проведено аналіз особливостей сучасної футбольної аналітики, підходів до використання event- та tracking-даних, моделей expected goals, а також сучасних методів машинного навчання, що застосовуються для прогнозування результативності ударів у футболі.

У результаті аналізу предметної галузі встановлено, що результативність ударів залежить від значної кількості просторових, часових та контекстних характеристик ігрового епізоду, а традиційні статистичні підходи мають обмежені можливості щодо врахування складних взаємозв'язків між цими параметрами. Це підтвердило доцільність використання методів машинного навчання для розв'язання задач прогнозування результативних дій.

У межах практичної частини роботи сформовано набір даних на основі відкритих футбольних даних StatsBomb. Для дослідження було використано 1296 матчів провідних футбольних турнірів, з яких сформовано датасет, що містить 32736 ударів по воротах. Виконано підготовку даних, формування цільової змінної, розрахунок додаткових просторових характеристик, зокрема дистанції до воріт та кута удару, а також проведено дослідницький аналіз даних.

Для прогнозування результативності ударів реалізовано та досліджено п'ять моделей машинного навчання: Logistic Regression, Random Forest, XGBoost, CatBoost та багатошаровий перцептрон. Проведено їх навчання, тестування та комплексне оцінювання із використанням метрик Accuracy, Precision, Recall, F1-score, ROC-AUC, Log Loss та Brier Score.

За результатами експериментального дослідження встановлено, що найкращу якість прогнозування забезпечила модель CatBoost, яка продемонструвала найвищі значення F1-міри для класу результативних ударів, найбільше значення ROC-AUC та найкращі показники якості ймовірнісних прогнозів. Отримані результати підтвердили перевагу ансамблевих градієнтних моделей над базовими лінійними та нейромережевими підходами для сформованого набору футбольних даних.

На основі отриманої моделі розроблено інтелектуальну систему прогнозування результативності ударів із використанням Streamlit. Реалізований вебінтерфейс забезпечує можливість аналізу реальних ударів із датасету, моделювання нових ігрових ситуацій, автоматичного розрахунку характеристик удару та отримання прогнозованої ймовірності голу. Проведене тестування показало коректну реакцію системи на зміну просторових характеристик ударів та підтвердило її практичну придатність для задач футбольної аналітики.

Отримані результати можуть бути використані для аналізу атакуювальних дій команд, оцінювання якості створених моментів, підтримки прийняття рішень у спортивній аналітиці та подальших досліджень у галузі прогнозування результативних дій на основі даних футбольних матчів.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Data-Driven football: how analytics is revolutionizing the sport. *4pballer*. URL: <https://4pballer.com/data-driven-football-how-analytics-is-revolutionizing-the-sport/> (date of access: 08.06.2026).
2. Football annotation: the hidden engine behind smarter game analysis. *pixelannotation.com*. URL: <https://pixelannotation.com/football-annotation-the-hidden-engine-behind-smarter-game-analysis/> (date of access: 08.06.2026).
3. Football analytics: interpreting event data with xG, xT and cutbacks. *footballytics*. URL: <https://www.footballytics.ch/post/data-analytics-practice-interpreting-event-data> (date of access: 08.06.2026).
4. Technical, tactical and spatial indicators related to goal scoring in European elite soccer / J. González-Ródenas et al. *Journal of human sport and exercise*. 2019. Vol. 15, no. 1. URL: <https://doi.org/10.14198/jhse.2020.151.17> (date of access: 08.06.2026).
5. How to: create A football shot map in R. *The False Nine | Logan | Substack*. URL: <https://thefalsenine.substack.com/p/how-to-create-a-football-shot-map> (date of access: 08.06.2026).
6. Anzer G., Bauer P. A goal scoring probability model for shots based on synchronized positional and event data in football. *Frontiers in sports and active living*. 2021. Vol. 3. URL: <https://doi.org/10.3389/fspor.2021.624475> (date of access: 09.06.2026).
7. Football visualization. *Barça Innovation Hub*. URL: <https://barcainnovationhub.fcbarcelona.com/blog/football-visualisation-capturing-chaos-and-cultivating-context/> (date of access: 09.06.2026).
8. Wisdom C., Javed A. Machine learning for data analytics in football: quantifying performance and enhancing strategic decision-making. *SSRN*. 2024. URL: <https://doi.org/10.2139/ssrn.4558733> (date of access: 09.06.2026).
9. M. Merzah B., S. Croock M., N. Rashid A. Intelligent classifiers for football player performance based on machine learning models. *International*

journal of electrical and computer engineering systems. 2024. Vol. 15, no. 2. P. 173–183. URL: <https://doi.org/10.32985/ijeces.15.2.6> (date of access: 09.06.2026).

10. Eryarsoy E., Delen D. Predicting the outcome of a football game: a comparative analysis of single and ensemble analytics methods. *Hawaii international conference on system sciences*. 2019. URL: <https://doi.org/10.24251/hicss.2019.136> (date of access: 10.06.2026).

11. Filiz E. Evaluation of match results of five successful football clubs with ensemble learning algorithms. *Research quarterly for exercise and sport*. 2022. P. 1–10. URL: <https://doi.org/10.1080/02701367.2022.2053647> (date of access: 10.06.2026).

12. Mead J., O'Hare A., McMenemy P. Expected goals in football: improving model performance and demonstrating value. *Plos one*. 2023. Vol. 18, no. 4. P. e0282295. URL: <https://doi.org/10.1371/journal.pone.0282295> (date of access: 10.06.2026).

13. Automatic event detection in football using tracking data / F. Vidal-Codina et al. *Sports engineering*. 2022. Vol. 25, no. 1. URL: <https://doi.org/10.1007/s12283-022-00381-6> (date of access: 10.06.2026).

14. Football event classification using convolutional autoencoder and multi-layer extreme learning machine / M. F. Hashmi et al. *IEEE Sensors Letters*. 2022. P. 1–4. URL: <https://doi.org/10.1109/lens.2022.3209366> (date of access: 10.06.2026).

15. Huang W., Wang S., Li P. The application of deep learning in tactical analysis of football matches. *Research Square*. 2025. URL: <https://doi.org/10.21203/rs.3.rs-7258508/v1> (date of access: 10.06.2026).

16. FootApp: an AI-powered system for football match annotation / S. Barra et al. *Multimedia tools and applications*. 2022. URL: <https://doi.org/10.1007/s11042-022-13359-0> (date of access: 10.06.2026).

17. OpenSTARLab: open approach for spatio-temporal agent data analysis in soccer / C. Yeung et al. *Complex & intelligent systems*. 2025. Vol. 11, no. 8. URL: <https://doi.org/10.1007/s40747-025-01965-y> (date of access: 10.06.2026).

18. Descriptive statistics: definition, formulas, types, examples. *GeeksforGeeks*. URL: <https://www.geeksforgeeks.org/maths/descriptive-statistics/> (date of access: 11.06.2026).

19. Attacking key performance indicators in soccer: current practice and perceptions from the elite to youth academy level / M. Herold et al. *Journal of sports science and medicine*. 2021. P. 158–169. URL: <https://doi.org/10.52082/jssm.2021.158> (date of access: 11.06.2026).

20. A new paradigm to understand success in professional football: analysis of match statistics in LaLiga for 8 complete seasons / D. Brito Souza et al. *International journal of performance analysis in sport*. 2019. Vol. 19, no. 4. P. 543–555. URL: <https://doi.org/10.1080/24748668.2019.1632580> (date of access: 11.06.2026).

21. Joshi M. A comparison between linear and logistic regression. *Medium*. URL: <https://medium.com/@maithilijoshi6/a-comparison-between-linear-and-logistic-regression-8aea40867e2d> (date of access: 11.06.2026).

22. Poisson distribution for football predictions: a mathematical guide. *Football Bet Prediction: Winning Strategies*. URL: <https://football-bet-prediction.com/football-predictions/> (date of access: 11.06.2026).

23. What is xG? 7 things expected goals reveals about football. *KharaSports*. URL: <https://kharasportsdaily.com/what-is-xg/> (date of access: 11.06.2026).

24. Kanade V. All you need to know about support vector machines. *Spiceworks*. URL: <https://www.spiceworks.com/soft-tech/what-is-support-vector-machine/> (date of access: 11.06.2026).

25. Yehoshua D. R. Random forests. *Medium*. URL: <https://medium.com/@roiyeo/random-forests-98892261dc49> (date of access: 11.06.2026).

26. Bagging vs boosting: what is the difference?. *Liora*. URL: <https://liora.io/en/bagging-vs-boosting> (date of access: 11.06.2026).
27. XGBoost & LGBM for time series forecasting: how to – 365 data science. 365 *Data Science*. URL: <https://365datascience.com/tutorials/python-tutorials/xgboost-lgbm/> (date of access: 11.06.2026).
28. Mead J., O'Hare A., McMenemy P. Expected goals in football: improving model performance and demonstrating value. *Plos one*. 2023. Vol. 18, no. 4. P. e0282295. URL: <https://doi.org/10.1371/journal.pone.0282295> (date of access: 12.06.2026).
29. Context is key: normalization as a novel approach to sport specific preprocessing of KPI's for match analysis in soccer / A. A. Phatak et al. *Scientific reports*. 2022. Vol. 12, no. 1. URL: <https://doi.org/10.1038/s41598-022-05089-y> (date of access: 12.06.2026).
30. Raabe D., Nabben R., Memmert D. Graph representations for the analysis of multi-agent spatiotemporal sports data. *Applied intelligence*. 2022. URL: <https://doi.org/10.1007/s10489-022-03631-z> (date of access: 12.06.2026).
31. Wisdom C., Javed A. Machine learning for data analytics in football: quantifying performance and enhancing strategic decision-making. *SSRN electronic journal*. 2023. URL: <https://doi.org/10.2139/ssrn.4558733> (date of access: 12.06.2026).
32. Understanding the confusion matrix in machine learning. *GeeksforGeeks*. URL: <https://www.geeksforgeeks.org/machine-learning/confusion-matrix-machine-learning/> (date of access: 12.06.2026).
33. Verronneau R., Cadiou B., Perchet V. EPS : an explainable post-shot expected goal metric for evaluating goalkeepers and attackers. *HAL open science*. 2025. URL: <https://hal.science/hal-05258651v1> (date of access: 12.06.2026).
34. Statsbomb/open-data: free football data from StatsBomb. *GitHub*. URL: <https://github.com/statsbomb/open-data> (date of access: 13.06.2026).