

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 моделі машинного навчання для прогнозування виникнення та
поширення лісових пожеж на основі супутникових даних та метеорологічних
показників
(тема)

Виконав:
здобувач _____ четвертого _____ року навчання,
групи _____ ІТШ-22-1
Дмитро УСІЧЕНКО
(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 _____ доц. Олександра ВІТЬКО
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ

_____ Лариса ЧАЛА
(підпис) (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 _____ Усіченку Дмитру Сергійовичу _____
(прізвище, ім'я, по батькові)

1. Тема роботи _____ Розробка моделі машинного навчання для прогнозування виникнення та поширення лісових пожеж на основі супутникових даних та метеорологічних показників _____

затверджена наказом університету від 5 червня 2026 р. № 488Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії 9 липня 2026 р.

3. Вихідні дані до роботи ретроспективні масиви метеорологічних спостережень та термальних аномалій; супутникові вегетаційні індекси (NDVI, NDWI) з відкритих баз NASA; науково-технічна література з екологічного моделювання та обчислювального інтелекту; інструментарій мови Python для машинного навчання (XGBoost, Scikit-learn) та математичні алгоритми інтерпретації моделей SHAP.

4. Перелік питань, що потрібно опрацювати в роботі _____

1) Аналіз предметної галузі та постановка задачі дослідження

2) Теоретичні засади машинного навчання та пояснювального штучного інтелекту

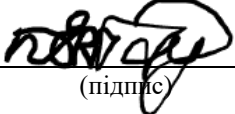
3) Математичне моделювання та методологія прогнозування ризику пожежі

4) Реалізація та аналіз результатів функціонування системи

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 наявних систем моніторингу	10.06.2026	виконано
3	Збір метео- та геоданих	13.06.2026	виконано
4	Проектування архітектури системи та підготовка датасетів	14.06.2026	виконано
5	Конструювання просторово-часових ознак	18.06.2026	виконано
6	Навчання та оптимізація моделей класифікації	20.06.2026	виконано
7	Оцінювання якості моделей та вибір фінального алгоритму	22.06.2026	виконано
8	Інтеграція модуля локального пояснення прогнозів (SHAP)	25.06.2026	виконано
9	Реалізація та тестування інтерактивного вебзастосунку	28.06.2026	виконано
10	Написання пояснювальної записки	30.06.2026	виконано
11	Перевірка на академічний плагіат	02.07.2026	виконано
12	Нормоконтроль	04.07.2026	виконано
13	Підготовка доповіді та презентації	05.07.2026	виконано
14	Попередній захист та рецензування	07.07.2026	виконано
15	Захист перед ЕК	09.07.2026	

Дата видачі завдання 8 червня 2026 р.

Здобувач 
(підпис)

Керівник роботи _____
(підпис)

доц. Олександра ВІТЬКО
(посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 98 с., 38 рис., 14 табл., 1 дод., 20 джерел.

АЛЖИР, БІНАРНА КЛАСИФІКАЦІЯ, ГЕОПРОСТОРОВИЙ АНАЛІЗ, ЛІСОВІ ПОЖЕЖІ, МАРОККО, МАШИННЕ НАВЧАННЯ, ПОЯСНЮВАЛЬНИЙ ШТУЧНИЙ ІНТЕЛЕКТ, СУПУТНИКОВІ ДАНІ, NDVI, RANDOM FOREST, SHAP, STREAMLIT, XGBOOST.

Об'єктом дослідження є процеси виникнення та поширення лісових пожеж під впливом метеорологічних, топографічних і геопросторових факторів у регіонах Північної Африки.

Предметом дослідження є алгоритми машинного навчання для аналізу геопросторових і метеоданих, методи пояснювального штучного інтелекту для інтерпретації прогнозів та технології розробки аналітичних вебзастосунків.

Метою роботи є розробка інтелектуальної інформаційної системи Forest Fire Intelligence System (FFIS) для прогнозування ризику лісових пожеж на основі супутникових та метеорологічних даних із забезпеченням інтерпретованості результатів та зручного вебінтерфейсу.

У роботі розроблено інформаційну систему FFIS для прогнозування ризику лісових пожеж у Марокко та Алжирі, реалізовану як повнофункціональний вебзастосунок. На основному датасеті (Morocco Wildfire Prediction) порівняно чотири класифікатори, серед яких найвищу точність продемонстрував алгоритм XGBoost. Для даних Алжиру моделі навчено на компонентах системи FWI. Інтегровано модуль TreeSHAP для глобального й локального пояснення прогнозів. Застосунок підтримує прогнозування в реальному часі, інтерактивну карту з шарами NASA GIBS, порівняння моделей, розвідувальний аналіз, сценарну симуляцію та генерацію AI-рекомендацій на основі локальної мовної моделі.

ABSTRACT

Bachelor's thesis contains: 98 pp., 38 fig., 34 tabl., 1 ann., 20 references.

ALGERIA, BINARY CLASSIFICATION, EXPLANATORY ARTIFICIAL INTELLIGENCE, FOREST FIRES, GEOSPATIAL ANALYSIS, MACHINE LEARNING, MOROCCO, NDVI, RANDOM FOREST, SATELLITE DATA, SHAP, STREAMLIT, XGBOOST.

The subject of the study is the processes of the onset and spread of forest fires under the influence of meteorological, topographical and geospatial factors in regions of North Africa.

The subject of the study comprises machine learning algorithms for analysing geospatial and meteorological data, explainable artificial intelligence methods for interpreting forecasts, and technologies for developing analytical web applications.

The aim of the work is to develop the Forest Fire Intelligence System for forecasting the risk of forest fires based on satellite and meteorological data, ensuring that the results are interpretable and that the system features a user-friendly web interface.

This study presents the FFIS information system for forecasting the risk of forest fires in Morocco and Algeria, implemented as a fully functional web application. On the main dataset (Morocco Wildfire Prediction), four classifiers were compared, with the XGBoost algorithm demonstrating the highest accuracy. For the Algerian data, the models were trained using components of the FWI system. The TreeSHAP module has been integrated to provide global and local explanations of the forecasts. The application supports real-time forecasting, an interactive map with NASA GIBS layers, model comparison, exploratory analysis, scenario simulation and the generation of AI recommendations based on a local language model.

ЗМІСТ

Вступ.....	9
1 Аналіз предметної галузі та постановка задачі дослідження	11
1.1 Опис предметної галузі	11
1.1.1 Кліматичні зміни та ключові фактори пожежної небезпеки	12
1.1.2 Еволюція методів прогнозування: від емпіричних моделей до машинного навчання.....	15
1.1.3 Роль супутникових даних у сучасному моніторингу	17
1.2 Проблеми та виклики в задачах прогнозування	20
1.3 Огляд та критичний аналіз існуючих рішень	24
1.4 Інноваційність та наукове підґрунтя розробленої системи	27
1.5 Об'єкти дослідження: регіони Марокко та Алжиру як репрезентативні території Північної Африки	30
1.5.1 Обґрунтування вибору географічного регіону дослідження.....	30
1.5.2 Фізико-географічна характеристика Марокко як основного об'єкта моделювання і Алжиру як додаткового об'єкта валідації. ...	31
1.6 Постановка задачі дослідження.....	32
2 Теоретичні засади машинного навчання та пояснювального штучного інтелекту.....	35
2.1 Огляд алгоритмів машинного навчання	35
2.1.1 Алгоритми бінарної класифікації.....	35
2.1.2 Теоретичні основи оцінювання якості класифікаційних моделей.	42
2.2 Теоретичні основи методів пояснювального штучного інтелекту	44
2.2.1 Проблема «чорного ящика» та метод SHAP.....	44
2.2.2 Алгоритм TreeSHAP для дерево-базованих моделей.....	47
2.3 Теоретичні основи аналізу геопросторових та супутникових даних	48
2.3.1 Принципи розрахунку спектральних вегетаційних індексів.....	48

2.3.2	Теоретичні засади часових лагових ознак та кумулятивних метеорологічних показників.	50
2.3.3	Геопросторова інтерполяція та просторова прив'язка даних. ..	51
2.4	Огляд сучасного технологічного інструментарію для розробки інтелектуальних систем аналізу даних.	52
2.5	Порівняльний теоретичний аналіз розглянутих підходів.	54
3	Математичне моделювання та методологія прогнозування ризику пожежі	57
3.1	Формальна постановка задачі прогнозування.....	57
3.2	Математичний апарат алгоритмів класифікації.....	59
3.2.1	Лінійні та ядрові моделі	59
3.2.2	Ансамблеві моделі	61
3.3	Математичний апарат методу SHAP.....	63
3.4	Формули супутникових індексів та інженерних ознак	65
3.5	Формули метрик оцінювання якості моделей.....	66
3.6	Формальна архітектура системи.....	66
4	Реалізація та аналіз результатів функціонування системи	73
4.1	Технологічний стек реалізації.....	73
4.2	Структура датасетів та опис вхідних даних	74
4.3	Реалізація пайплайну передобробки даних	77
4.4	Навчання та порівняльне оцінювання моделей	78
4.4.1	Результати навчання моделей для регіону Марокко.....	79
4.4.2	Результати навчання моделей для регіону Алжир	80
4.5	Реалізація модуля пояснювального ШП	82
4.5.1	Глобальна важливість ознак та інтерпретація (Марокко)	82
4.5.2	Глобальна важливість ознак (Алжир).....	84
4.5.3	Локальне пояснення (waterfall).....	85
4.6	Опис реалізованого вебзастосунку.....	85
4.6.1	Загальна структура інтерфейсу.....	85
4.6.2	Модуль прогнозування в реальному часі	86

4.6.3	Геопросторова карта	87
4.6.4	Модуль AI-рекомендацій	88
4.6.5	Модуль сценарної симуляції та звітності.....	89
4.7	Тестування системи та обговорення результатів.....	91
4.7.1	Тестування коректності прогнозування та обмеження системи.....	91
4.7.2	Порівняльний аналіз двох регіональних моделей	92
	Висновки	93
	Перелік джерел посилання	95
	Додаток А Відомість кваліфікаційної роботи	98

ВСТУП

Зростання кількості неконтрольованих лісових пожеж через глобальні кліматичні зміни становить критичну загрозу для природних екосистем. Особливо гостро ця проблема постає в посушливих регіонах Північної Африки (Марокко, Алжир), де екстремальні температури та мінімальна кількість опадів створюють ідеальні умови для займань. Традиційні системи моніторингу, що базуються на емпіричних метеорологічних індексах, виявляються малоефективними, оскільки не здатні враховувати складні нелінійні взаємозв'язки між природними факторами. Перехід до інтелектуального аналізу просторових даних на базі машинного навчання є необхідним кроком для забезпечення проактивного прогнозування зон ризику.

Мета роботи – розробка інтелектуальної інформаційної системи прогнозування виникнення та поширення лісових пожеж шляхом інтеграції супутникових та метеорологічних даних. Для досягнення мети вирішено такі завдання:

- провести аналіз предметної галузі та критичний огляд існуючих систем моніторингу;
- здійснити препроцесинг гетерогенних даних регіонів Північної Африки;
- дослідити та навчити ансамблеві моделі (XGBoost, Random Forest) для класифікації рівня пожежної небезпеки;
- застосувати методи пояснювального ШІ (SHAP) для інтерпретації логіки моделі;
- розробити інтерактивний дашборд для візуалізації прогнозів на карті.

Об'єкт дослідження – процеси виникнення та поширення лісових пожеж.

Предмет дослідження – алгоритми машинного навчання та програмні інструменти інтелектуальної обробки геопросторових і метеорологічних даних для прогнозування ризику займання.

Методи дослідження: системний аналіз (дослідження предметної галузі), математична статистика (обробка предикторів), методи машинного навчання на основі дерев рішень (побудова класифікаторів) та теорія ігор (SHAP для інтерпретації результатів).

Наукова новизна полягає у використанні гібридного підходу, який поєднує високу точність алгоритмів градієнтного бустингу (XGBoost) з методами пояснювального інтелекту. Це дозволяє не лише прогнозувати рівень небезпеки, а й кількісно оцінювати вплив кожного конкретного кліматичного фактора на результат.

Практичне значення полягає у створенні діючого прототипу системи підтримки прийняття рішень для екологічних служб, що забезпечує оперативний моніторинг ризиків та легко масштабується на нові географічні регіони.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Опис предметної галузі

Проблема своєчасного виявлення, моніторингу та прогнозування лісових пожеж є однією з найскладніших і найважливіших завдань сучасної прикладної екології та геоінформатики. У глобальному вимірі лісові пожежі становлять масштабну загрозу, що виходить далеко за межі локального знищення деревини. Згідно з дослідженнями міжнародних екологічних інституцій, неконтрольоване поширення вогню призводить до комплексного руйнування біоценозів: відбувається глибока деградація та ерозія ґрунтів, порушується гідрологічний цикл цілих регіонів, знищуються природні середовища існування тисяч видів флори та фауни.

Окрім прямої шкоди екосистемам, масштабні лісові пожежі відіграють критичну роль у зміні хімічного складу атмосфери. Величезні обсяги діоксиду вуглецю (CO_2), метану, оксидів азоту та дисперсних аерозолів, що вивільняються під час інтенсивного горіння органічної біомаси, потрапляють у верхні шари тропосфери. Це явище створює ефект позитивного кліматичного зворотного зв'язку: викиди парникових газів від пожеж прискорюють темпи глобального потепління, що, своєю чергою, формує ще більш сприятливі кліматичні умови для виникнення нових, інтенсивніших загорянь у майбутньому.

Зважаючи на зростаючу небезпеку та катастрофічні наслідки неконтрольованого поширення вогню, розуміння природи цього явища вимагає глибокого міждисциплінарного підходу. Сучасна наука розглядає лісову пожежу не як випадкове явище, а як складний фізико-хімічний процес, що суворо визначається станом навколишнього середовища. Довгий час превентивні заходи обмежувалися лише візуальним патрулюванням та констатацією факту займання. Проте сьогодні, в епоху цифровізації та

накопичення масивів великих даних (Big Data), фокус предметної галузі зміщується з реактивного реагування на розробку інтелектуальних предиктивних (прогностичних) систем. Такі системи покликані завчасно ідентифікувати зони критичного ризику на основі аналізу сотень терабайтів метеорологічних та геопросторових даних.

1.1.1 Кліматичні зміни та ключові фактори пожежної небезпеки

Процес виникнення, розвитку та просторового поширення лісової пожежі чітко підпорядковується законам термодинаміки та екології. Фундаментальною основою для моделювання пожежної небезпеки є концепція «пожежного трикутника навколишнього середовища» (Fire Environment Triangle). Ця концепція постулює, що поведінка вогню залежить від безперервної та одночасної взаємодії трьох ключових груп предикторів (рисунок 1.1):

- метеорологічні фактори. Це найбільш динамічна складова системи. До неї належать температура повітря, відносна вологість, швидкість та напрямок вітру, а також рівень інсоляції. У сучасних прогностичних моделях особлива увага приділяється дефіциту тиску водяної пари (VPD – Vapor Pressure Deficit), який визначає здатність атмосфери «витягувати» вологу з рослинності. Вітер не лише постачає кисень у зону горіння, але й фізично нахиляє полум'я, прискорюючи конвективне нагрівання сусідніх дерев;

- топографічні (орографічні) фактори. Геоморфологія місцевості – рельєф, кут нахилу схилу, висота над рівнем моря та експозиція – відіграє роль просторового каталізатора. Відомо, що фронт полум'я поширюється вгору по стрімкому схилу в геометричній прогресії швидше, ніж на рівнині. Крім того, південні та західні схили (у Північній півкулі) отримують більше сонячної радіації, що призводить до швидшого висихання ґрунту та рослинності;

– характеристики пального матеріалу (Fuel Load). Це обсяг, структура та хімічний склад доступної біомаси. Враховуються як жива вологість пального (LFMC – Live Fuel Moisture Content), так і вологість мертвої деревини (DFMC). Структура лісового намету, наявність густого чагарника та хвойної підстилки визначають інтенсивність горіння та ймовірність переходу низової пожежі у верхову [19].

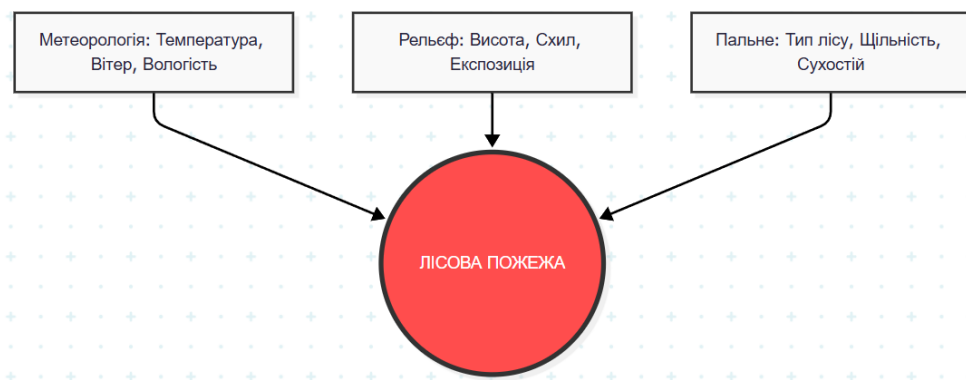


Рисунок 1.1 – Візуалізація «Макропожежного трикутника навколишнього середовища» (Fire Environment Triangle)

Зазначені три групи факторів не діють ізольовано – саме їхня синергетична взаємодія визначає кінцевий рівень пожежної небезпеки для конкретної території в конкретний момент часу. Наприклад, високий об'єм накопиченого пального матеріалу стає критично небезпечним лише за умови одночасного впливу низької атмосферної вологості та сильного вітру, тоді як кожен із цих факторів окремо може не призвести до займання. Саме ця мультиплікативна природа пожежного ризику зумовлює необхідність використання методів машинного навчання, здатних автоматично виявляти подібні багатовимірні залежності у великих масивах даних.

Особливої гостроти розуміння цих факторів набуло на тлі сучасних глобальних кліматичних змін. Систематичне зростання температурних оптимумів і зміщення кліматичних поясів призводять до виникнення безпрецедентних періодів посухи. У межах даного дослідження цільовим

регіоном для побудови моделей машинного навчання виступає територія Північної Африки, зокрема гірські та передгірні масиви Королівства Марокко та Алжирської Народної Демократичної Республіки [18].

Цей макрорегіон належить до середземноморського кліматичного поясу і характеризується екстремально спекотним та посушливим літом. Кліматична вразливість цієї території зумовлена її географічним розташуванням між Атлантичним океаном, Середземним морем та найбільшою пустелею світу – Сахараю. Головним метеорологічним каталізатором пожежної небезпеки тут виступають місцеві пустельні вітри: Шергі (Chergui) у Марокко та Сироко (Sirocco) в Алжирі. Ці вітри приносять маси розпеченого сухого повітря з Сахари. Долаючи гірські хребти Атласу, повітряні маси зазнають адіабатичного нагрівання (феновий ефект). У результаті температура повітря в долинах та на узбережжі може за лічені години сягнути 45–48 °C, а відносна вологість впасти до критичних 5–10%.

Такий екстремальний мікроклімат накладається на специфічну рослинність регіону. Гірські масиви Середнього та Високого Атласу вкриті ендемічними хвойними лісами, де домінує сосна алеппська (*Pinus halepensis*) та атласький кедр. Хвоя та кора цих дерев перенасичені легкозаймистими ефірними оліями (терпенами) та смолами, які при високих температурах починають активно випаровуватися, створюючи над лісом вибухонебезпечну газову суміш. Підлісок у цих регіонах складається з густих, майже непрохідних заростей ксерофітних чагарників – маквісу та гариги. Під час літньої посухи чагарники втрачають значну частину вологи, перетворюючись на ідеальне пальне високої щільності.

Взаємодія розпеченого пустельного вітру з пересушеними смолистими лісами Атласу створює умови, за яких будь-яка іскра миттєво переростає в катастрофічну верхову пожежу. Саме тому Північна Африка є одним із найскладніших, але водночас найбільш репрезентативних регіонів для розробки та валідації інтелектуальних моделей прогнозування пожежних ризиків на базі супутникових і метеорологічних даних.

1.1.2 Еволюція методів прогнозування: від емпіричних моделей до машинного навчання

Історичний ретроспективний аналіз розвитку систем екологічної безпеки та лісового менеджменту свідчить, що методологія оцінювання пожежної небезпеки пройшла тривалий і складний шлях еволюції. Десятиліттями традиційні методи прогнозування базувалися виключно на статистичних спостереженнях та жорстких детермінованих емпіричних моделях. Фундаментальною основою таких моделей була побудова систем математичних індексів, які намагалися агрегувати декілька ключових метеорологічних показників в єдине числове значення – рівень ризику.

Найбільш відомими та широко використовуваними прикладами такого класичного підходу є Канадський індекс пожежної небезпеки лісів – Fire Weather Index, Національна система оцінки небезпеки пожеж США, а також класичний індекс Нестерова, розроблений ще в середині XX століття. Незважаючи на їхню історичну значущість та відносну простоту обчислення, сучасна наука визнає, що ці класичні підходи мають низку критичних математичних та просторових обмежень:

- просторова екстраполяція низької точності: емпіричні індекси розраховуються на основі даних, отриманих із розрізнених наземних метеорологічних станцій. У гірських регіонах (таких як гори Атлас) відстань між станціями може складати десятки або сотні кілометрів. Класичні системи застосовують метод лінійної екстраполяції цих точкових даних на величезні площі, повністю ігноруючи складну топографію та локальні мікрокліматичні аномалії;

- математична ригідність (статичність формул): традиційні індекси використовують наперед задані, жорсткі математичні коефіцієнти. Вони не здатні адаптуватися до нестандартних ситуацій та ефективно виявляти складні, нелінійні і мультиплікативні взаємозв'язки між природними факторами. Наприклад, класична формула не може врахувати ситуацію,

коли ефект від підвищення швидкості вітру повністю нівелюється раптовою зміною типу підстильної рослинності;

– відсутність здатності до навчання: емпірична модель завжди видаватиме однаковий результат при однакових вхідних параметрах, незалежно від того, чи підтвердився її попередній прогноз у реальності. Вона не акумулює історичний досвід.

Зважаючи на вказані фундаментальні недоліки, в умовах сучасного стрімкого зростання обсягів геопросторових даних (Big Data), наукова спільнота та інженери-практики здійснюють глобальний перехід до парадигми інтелектуального моніторингу. Ця нова парадигма повністю базується на застосуванні сучасних алгоритмів машинного навчання.

Впровадження методів машинного навчання докорінно змінює методологію моделювання природних катастроф. На відміну від статичних емпіричних формул, алгоритми ML є системами, що керуються даними. Вони здатні самостійно ідентифікувати приховані багатовимірні патерни у масивах історичних даних, знаходячи неочевидні кореляції між супутниковою телеметрією, метеорологічними показниками та фактичними координатами займань (рисунок 1.2).

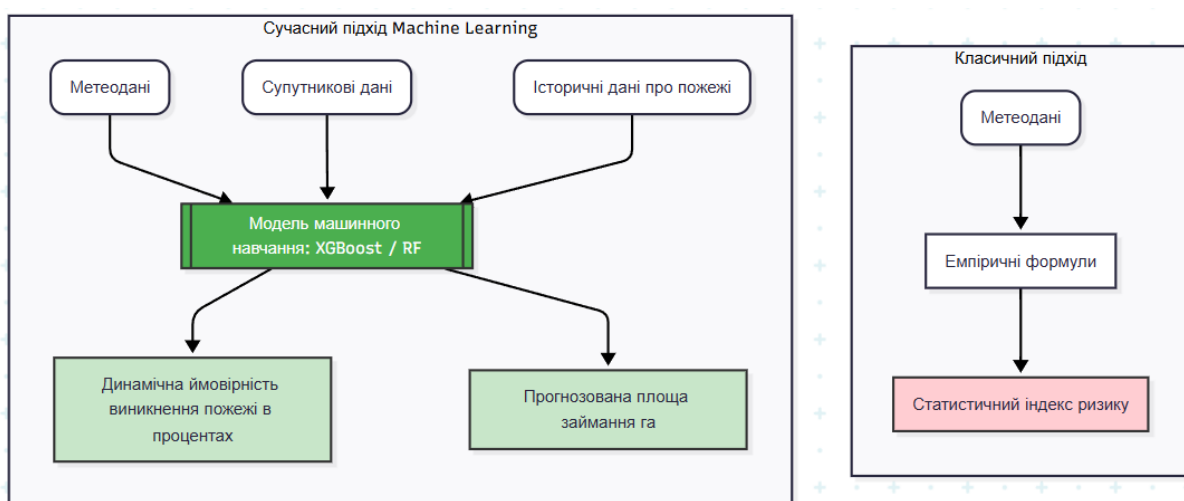


Рисунок 1.2 – Порівняльна блок-схема класичного підходу проти ML

Використання таких підходів, зокрема ансамблевих дерев рішень та нейронних мереж, дозволяє:

- обробляти гетерогенні дані (одночасно аналізувати числові метеодані та растрові супутникові знімки);
- досягати безпрецедентно високої просторової роздільної здатності – розраховувати ризик не для цілого регіону, а для конкретного квадрата площею 1х1 км або навіть 375х375 метрів;
- трансформувати систему екологічного моніторингу з реактивної, тобто констатації факту наявності термальної аномалії, на предиктивну, що визначає точну зону майбутнього ризику за кілька днів до моменту потенційного займання.

Окрім зазначених переваг, застосування ML-підходів забезпечує можливість безперервного покращення моделей у міру надходження нових даних спостережень. На відміну від статичних емпіричних формул, що потребують повної переробки при зміні кліматичних умов, моделі машинного навчання можуть бути додатково навчені на актуальних даних із мінімальними витратами обчислювальних ресурсів, що забезпечує їхню адаптивність до довгострокових кліматичних трендів.

1.1.3 Роль супутникових даних у сучасному моніторингу

Невід'ємною складовою та головним джерелом вхідної інформації для сучасних предиктивних ML-систем є використання даних дистанційного зондування Землі (ДЗЗ) [10]. Інтеграція аерокосмічних методів забезпечує глобальне покриття, безперервність спостережень та можливість моніторингу найбільш важкодоступних регіонів (таких як гірські ущелини та пустельні кордони). Найбільший та найякісніший масив відкритих геопросторових даних сьогодні генерується супутниковими угрупованнями EOS (Earth Observing System) Національного управління з аеронавтики і дослідження космічного простору США (рисунок 1.3).

Розуміння фізичних принципів роботи їхніх ключових мультиспектральних сенсорів – MODIS та VIIRS – є критично важливим для правильного формування векторів ознак (feature vectors) у машинному навчанні.

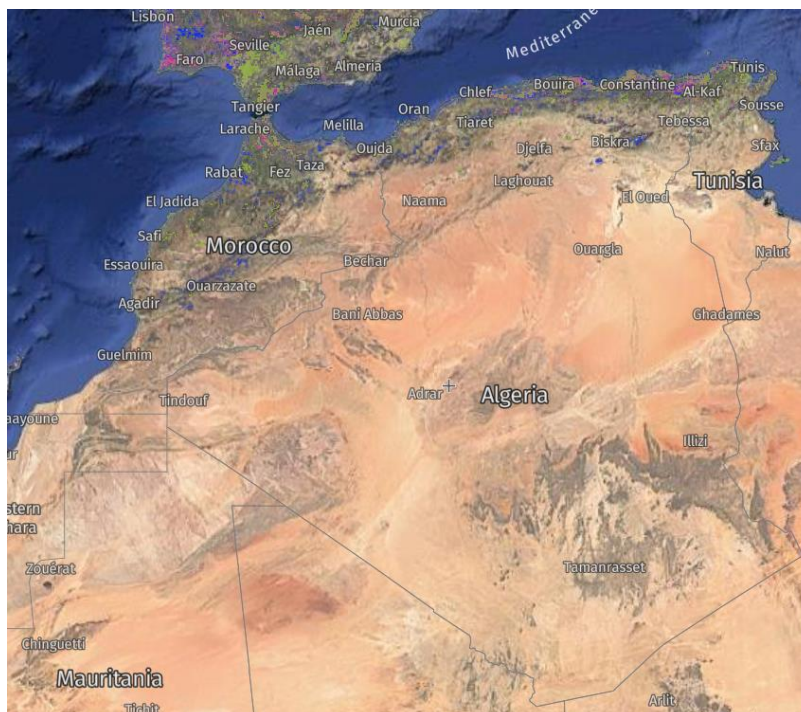


Рисунок 1.3 – Фізико-географічна характеристика та зони лісових масивів гір Атлас (Північна Африка). Сформовано на базі геоданих Global Forest Watch

Варто зазначити, що якість та повнота супутникових даних суттєво залежить від хмарного покриття, кута спостереження та часу проходження супутника над цільовою територією. Для регіону Північної Африки, що характеризується переважно ясною погодою в літні місяці, ці обмеження є менш критичними порівняно з тропічними або помірними широтами, що робить цей регіон особливо придатним для побудови ML-моделей на основі супутникових даних.

Сенсор MODIS (Moderate Resolution Imaging Spectroradiometer), розміщений на борту геліосинхронних супутників Terra та Aqua, є одним із найуспішніших інструментів в історії спостереження за Землею [7]. Він

проводить безперервну зйомку поверхні у 36 різних спектральних діапазонах електромагнітного випромінювання. Для безпосередньої ідентифікації активних осередків пожеж застосовується спеціалізований контекстний алгоритм MOD14/MYD14. Цей алгоритм аналізує радіаційну температуру в середньому інфрачервоному діапазоні (довжина хвилі ≈ 4 мкм) та порівнює її з фоновою температурою сусідніх «здорових» пікселів у тепловому інфрачервоному діапазоні (≈ 11 мкм) [7]. Завдяки такій різниці яскравості (Brightness Temperature Difference) сенсор MODIS здатний надійно фіксувати субпіксельні пожежі. Це означає, що при номінальній просторовій роздільній здатності у 1 кілометр система здатна виявити займання площею всього 50–100 квадратних метрів, що ідеально підходить для фіксації пожеж на ранніх стадіях.

Більш сучасним та технологічно інноваційним інструментом є сенсор VIIRS (Visible Infrared Imaging Radiometer Suite), який встановлений на метеорологічних супутниках нового покоління Suomi NPP та NOAA-20 [8]. VIIRS розроблений для продовження та покращення місії MODIS. Його головна перевага полягає у значно вищій просторовій роздільній здатності – 375 метрів на піксель у спеціальних I-каналах (Imagery bands) [8]. Таке підвищення деталізації мінімізує похибку геопозиціонування та дозволяє ML-моделям виявляти пожежі з нижчою температурою горіння (наприклад, тліючу лісову підстилку або торф'яники). Особливістю VIIRS є наявність унікального каналу DNB (Day/Night Band), який здатен з високою точністю картувати периметр поширення вогню навіть у нічний час.

Окрім прямої фіксації термальних аномалій (реєстрації цільової змінної для ML-моделі), мультиспектральні оптичні дані з цих сенсорів є критично важливим джерелом для розрахунку спектральних вегетаційних індексів. Фундаментальним предиктором у задачах екологічного моніторингу виступає нормалізований диференційний вегетаційний індекс – NDVI (Normalized Difference Vegetation Index) [6].

З точки зору фізики та біології, індекс NDVI базується на специфічних властивостях фотосинтезу. Молекули хлорофілу в здоровому листі рослин інтенсивно поглинають електромагнітне випромінювання у видимому червоному діапазоні спектра (для отримання енергії) та активно відбивають світло у ближньому інфрачервоному діапазоні (NIR), щоб уникнути перегріву тканин. Математично NDVI розраховується як відношення різниці цих двох каналів до їхньої суми.

У контексті розробки системи прогнозування пожежної небезпеки значення NDVI виконує функцію високонадійного проксі-індикатора вологозабезпеченості та фізіологічного стану лісового покриву. Високі значення індексу (ближче до 1.0) свідчать про густу, здорову та насичену вологою крону. Натомість, різке або поступове зниження значень NDVI у регіоні (до позначок 0.2 – 0.3) свідчить про глибокий водний стрес рослин, скидання листя та накопичення сухих, легкозаймистих матеріалів. Алгоритми машинного навчання (такі як XGBoost чи Random Forest) використовують цю просторову динаміку NDVI як один із найвагоміших маркерів для розрахунку ймовірності неконтрольованого займання на конкретній географічній ділянці.

1.2 Проблеми та виклики в задачах прогнозування

Процес розробки та імплементації інтелектуальних інформаційних систем для екологічного моніторингу є нетривіальною задачею, яка супроводжується низкою специфічних математичних, фізико-географічних та суто інженерних викликів. Глибокий аналіз цих проблемних аспектів є обов'язковим етапом дослідження, оскільки саме вони визначають стратегію вибору кінцевої архітектури програмного рішення, алгоритмів машинного навчання та методів попередньої обробки просторових даних.

Лісові пожежі, незважаючи на їхню масштабність та катастрофічну руйнівність, у загальному масиві екологічних спостережень є статистично

рідкісними аномаліями. У будь-якому репрезентативному наборі геопросторових даних кількість записів, що фіксують нормальний стан середовища (відсутність займання), потенційно може на декілька порядків перевищувати кількість зафіксованих пожеж (у необроблених джерелах це співвідношення нерідко сягає 1:100 або навіть 1:1000). При використанні стандартних алгоритмів машинного навчання такий дисбаланс неминуче призводить до феномену «парадоксу точності» (Accuracy Paradox). Математичній моделі під час процесу мінімізації функції втрат стає алгоритмічно вигідніше завжди генерувати прогноз «відсутність пожежі», досягаючи при цьому фіктивної загальної точності (Accuracy) на рівні 99%, попри те, що модель фактично не виявляє жодної реальної пожежі.

Для подолання цієї проблеми дослідники відмовляються від орієнтації виключно на класичну метрику Accuracy на користь комплексного аналізу метрик Precision, Recall, F1-міри та площі під кривою точності-повноти, а також застосовують спеціалізовані методи балансування – стратифіковану вибірку при формуванні навчального набору даних, методи зважування класів або методи штучного балансування (наприклад, SMOTE). Питання дисбалансу класів та обраний у даній роботі підхід до його вирішення детально розглядаються у розділі 3.

Критичною вразливістю систем, побудованих на основі супутникового зондування, є фізична залежність якості вхідних даних від поточних гідрометеорологічних умов. Оптичні мультиспектральні сенсори, наприклад MODIS та VIIRS, є «сліпими» за умови наявності щільного хмарного покриття – вони не здатні фіксувати теплове випромінювання поверхні Землі крізь атмосферні перешкоди. Аналогічна проблема виникає і під час самої пожежі: густі шари аерозольного диму екранують інфрачервоне випромінювання, спотворюючи показники температури поверхні та спектральних вегетаційних індексів.

Це призводить до виникнення значних масивів пропущених значень у датасетах саме в ті критичні моменти, коли лісова екосистема потребує

найбільш ретельного моніторингу [12]. Наземні метеорологічні станції, своєю чергою, також не позбавлені технічних збоїв датчиків та періодичних втрат зв'язку, що додатково ускладнює формування повноцінних, неперервних часових рядів. Вирішення цієї проблеми вимагає застосування спеціалізованих методів імпутації (заповнення) пропущених значень на етапі попередньої обробки даних, що детально описано в розділі 3.

Фізика виникнення та поширення лісової пожежі принципово не підпорядковується лінійним закономірностям. Наприклад, вплив швидкості вітру на інтенсивність горіння має складний, нерідко немонотонний характер: вітер стимулює розповсюдження вогню, постійно постачаючи кисень у зону реакції, проте при досягненні штормових значень або при раптовій зміні рельєфних перешкод динаміка фронту полум'я стає хаотичною та турбулентною. Вологість ґрунту взаємодіє з температурою також нелінійно – поєднання помірно високої температури з критично низькою вологістю може давати непропорційно більший приріст ризику, ніж сума окремих ефектів кожного фактора.

Класичні моделі множинної лінійної регресії виявляються математично неспроможними адекватно апроксимувати таку поведінку, що диктує необхідність застосування складних нелінійних моделей – зокрема, ансамблів на основі дерев рішень, які здатні самостійно ідентифікувати складні перехресні взаємодії (interactions) між десятками предикторів без необхідності їх явного попереднього програмування дослідником.

Сучасні предиктивні системи вимагають агрегації даних із різних, часто технологічно несумісних джерел. Метеорологічні показники, а саме вологість, швидкість вітру, температура, зазвичай збираються мережею стаціонарних наземних метеостанцій і представляють собою точкові просторові дані (Point Data) з високою частотою оновлення. Водночас супутникові індекси (наприклад, NDVI) формують безперервну растрову просторову сітку (Grid/Raster Data) з суттєво нижчою частотою оновлення.

Проблема коректного узгодження цих різнорідних просторових структур (Spatial mismatch) вимагає впровадження методів геопросторової інтерполяції та розрахунку часових лагів (lag-features) на етапі попередньої підготовки датасету – підходу, що дозволяє врахувати не лише миттєвий стан середовища, а й кумулятивну динаміку метеорологічних умов за попередній період (наприклад, 7–15 днів), що є визначальним фактором накопиченого дефіциту вологи в рослинності та ґрунті (таблиця 1.1).

Таблиця 1.1 – Зведена характеристика основних викликів предметної галузі та підходів до їх вирішення

№	Виклик	Сутність проблеми	Підхід до вирішення
1	Дисбаланс класів	Кількість спостережень «без пожежі» суттєво перевищує кількість спостережень «пожежа», що призводить до «парадоксу точності»	Стратифікована вибірка, метрики Precision/Recall/F1/AUC замість Accuracy, балансування класів
2	Атмосферні перешкоди	Хмарність та дим екранують супутникові сенсори, спричиняючи пропуски даних саме в критичні моменти	Методи імпутації пропущених значень (медіанна імпутація), використання багатоджерельних даних
3	Нелінійність взаємодій	Фактори ризику (вітер, вологість, температура) взаємодіють нелінійно та мультиплікативно	Застосування ансамблевих нелінійних моделей (Random Forest, XGBoost) замість лінійної регресії
4	Просторово-часова гетерогенність	Метеодані – точкові й часті, супутникові – растрові й рідші; потрібне узгодження	Розрахунок лагових ознак (lag-features), агрегація в часові вікна, просторова прив'язка до координат

Як свідчать дані таблиці, побудова надійної моделі вимагає подолання низки обмежень екологічних даних. Проблема дисбалансу класів вирішується застосуванням методів балансування та метрик F1-міри і ROC-AUC замість Accuracy. Фізичні перешкоди супутникового моніторингу нівелюються методами імпутації пропущених значень. Складна нелінійність взаємодії метеофакторів обґрунтовує використання

ансамблевих алгоритмів (XGBoost, Random Forest). Водночас просторово-часова гетерогенність джерел ефективно долається шляхом інженерії лагових ознак, що перетворює розрізнені спостереження на єдину узгоджену матрицю.

1.3 Огляд та критичний аналіз існуючих рішень

Світовий ринок геоінформаційних технологій та платформ екологічного моніторингу на сьогодні пропонує низку масштабних програмних рішень, спрямованих на відстеження лісових пожеж. Проте глибокий критичний аналіз їхньої системної архітектури та функціональних можливостей дозволяє виявити суттєвий дисбаланс між наявним інструментарієм та сучасними світовими вимогами до проактивного реагування на надзвичайні ситуації. Більшість існуючих рішень залишається заручниками застарілих математичних моделей або обмеженими виключно функцією констатації факту.

Беззаперечними лідерами в галузі макрорегіонального моніторингу виступають еталонні рішення, розроблені за підтримки структур Європейської Комісії [13]: Європейська інформаційна система лісових пожеж та Глобальна інформаційна система лісових пожеж [9], [15]. Головними безперечними перевагами цих платформ є глобальне охоплення територій та наявність фундаментальних, добре структурованих архівів ретроспективних даних за останні десятиліття. Незважаючи на це, архітектура таких систем характеризується надмірною технологічною громіздкістю, що зумовлює складність їхньої локальної інтеграції для потреб конкретних територіальних громад. Їхня базова функціональна парадигма зосереджена переважно на реактивному моніторингу, а саме візуалізації теплових аномалій та картування вигорілих площ. У свою чергу, їхні прогностичні модулі (Fire Danger Forecast) цілком і повністю базуються на класичних емпіричних метеорологічних індексах, зокрема на канадській

системі FWI. Використання статичних формул означає, що ці системи ігнорують можливості локального ансамблевого машинного навчання та не здатні адаптивно підлаштовуватися під приховані мікрокліматичні патерни унікальних регіонів, таких як гірські масиви Північної Африки.

Окремої наукової уваги заслуговує глобальна аерокосмічна система NASA FIRMS (Fire Information for Resource Management System) [14]. На сьогодні це один із найбільш оперативних та технологічно досконалих інструментів у світі, який надає дані про активні пожежі у режимі, максимально наближеному до реального часу (Near Real-Time, NRT). FIRMS агрегує інформацію з оптичних сенсорів MODIS та VIIRS із затримкою лише у 3–6 годин після прольоту супутника над територією, забезпечуючи відкритий доступ до архівних та оперативних даних через зручний картографічний вебінтерфейс і програмний інтерфейс (API) [14].

Проте критичним концептуальним недоліком платформи FIRMS є її виключно спостережний, дескриптивний характер. Система з надзвичайно високою точністю констатує факт наявності термальної аномалії, але робить це тоді, коли лісовій екосистемі вже завдається незворотна шкода. FIRMS не має вбудованих предиктивних модулів і не надає дослідникам чи рятувальникам інструментів для випереджального аналізу ризиків на основі синтезу метеорологічних та геопросторових предикторів. Фактично, FIRMS відповідає на питання «де зараз горить», але не відповідає на питання «де ймовірно загориться найближчим часом» – саме цю прогалину покликана заповнити система, що розробляється в межах даної роботи.

Паралельно на рівні окремих держав функціонують локальні системи оповіщення державних служб з надзвичайних ситуацій. Їхньою єдиною перевагою є використання верифікованих даних з мережі офіційних наземних метеостанцій у режимі реального часу. Проте прогнози від таких служб зазвичай характеризуються надзвичайно низькою просторовою роздільною здатністю – наприклад, встановлюється єдиний, генералізований рівень пожежної небезпеки для цілої адміністративної

області чи провінції площею у тисячі квадратних кілометрів. Це повністю унеможливило ефективний точковий розподіл ресурсів рятувальних підрозділів.

Крім того, результати моніторингу подаються у вигляді консервативних текстових або табличних звітів без використання сучасних методів інтерактивної картографії та прозорих аналітичних моделей, що ускладнює їх сприйняття та оперативне використання як фахівцями, так і населенням. Відсутність механізмів пояснення того, чому конкретній території присвоєно той чи інший клас небезпеки, додатково знижує довіру до подібних систем серед кінцевих користувачів.

Для наочного та структурованого зіставлення функціональних можливостей, а також технологічних обмежень існуючих платформ із розроблюваною інтелектуальною системою, було сформовано детальну порівняльну таблицю (таблиця 1.2).

Таблиця 1.2 – Порівняльний аналіз існуючих інформаційних систем моніторингу лісових пожеж

Критерій оцінювання	EFFIS / GWIS	NASA FIRMS	Локальні системи ДСНС	Запропонована система (FFIS)
Основна парадигма	Реактивно-статистична	Виключно реактивна (констатація факту)	Статистично-описова	Предиктивна (проактивний прогноз)
Алгоритми Machine Learning	Відсутні (емпіричні індекси)	Відсутні (трансляція даних сенсорів)	Відсутні	Наявні (ансамблі XGBoost, RF)
Просторова роздільна здатність	Середня / Низька (макрорегіональна)	Висока (до 375 м)	Дуже низька (на рівні області)	Висока (за координатною сіткою)
Джерела даних	Метеодані + Супутникові архіви	Супутникові оптичні сенсори	Наземні метеостанції	Синтез метеоданих та супутникових індексів
Пояснюваність (Explainable AI)	Відсутня	Відсутня	Відсутня	Наявна (модуль SHAP-аналітики)

Зіставлення виявлених концептуальних та інженерних недоліків дозволяє зробити обґрунтований висновок, що на поточному етапі розвитку геоінформаційних технологій спостерігається певний технологічний вакуум. Жодна з розглянутих систем не поєднує одночасно: проактивний, а не реактивний характер прогнозування; застосування адаптивних алгоритмів машинного навчання замість статичних емпіричних формул; високу просторову роздільну здатність на рівні конкретної географічної точки, а не цілого регіону чи провінції; наявність механізмів пояснення прийнятих моделлю рішень (Explainable AI). Існує нагальна науково-практична потреба у створенні спеціалізованих інформаційних систем, які б інтегрували високу просторову роздільну здатність прогнозів, отриманих за допомогою ансамблевих моделей машинного навчання, з сучасними модулями пояснювального штучного інтелекту.

1.4 Інноваційність та наукове підґрунтя розробленої системи

Зважаючи на виявлені критичні обмеження існуючих світових рішень, концептуальна архітектура розроблюваної інформаційної системи будується на інноваційному синтезі алгоритмів обчислювального інтелекту, методів геопросторового аналізу та технологій пояснювального машинного навчання (Explainable AI) [10]. Такий гібридний підхід дозволяє подолати технологічний вакуум, перетворюючи систему з інструменту пасивної констатації факту пожежі на потужну систему проактивної підтримки прийняття рішень (Decision Support System).

В основу розробленої системи покладено концепцію, яку можна означити як «Data-Driven Environmental Intelligence» – підхід, що передбачає цілковиту відмову від жорстко заданих емпіричних формул (на кшталт класичного індексу FWI чи індексу Нестерова) на користь гнучких статистичних моделей, здатних самостійно адаптуватися до специфіки

конкретного географічного регіону та навчатися безпосередньо на масивах історичних даних про реальні пожежні події.

Принципова відмінність цього підходу від класичних індексів полягає в тому, що замість експертного, апріорного встановлення ваги кожного окремого фактора (підхід, що неминуче є суб'єктивним і не здатний оптимально врахувати складні нелінійні взаємодії між десятками змінних одночасно), математична модель самостійно визначає оптимальні ваги та нелінійні залежності на основі статистичного аналізу сотень тисяч реальних спостережень. Це дозволяє моделі «відкривати» приховані закономірності – наприклад, специфічну для конкретного географічного регіону взаємодію між вологістю ґрунту, типом рослинності та локальними вітровими режимами – які жоден дослідник не зміг би апріорно закласти у фіксовану аналітичну формулу [11].

Ключовою практичною перевагою такого підходу для предметної галузі прогнозування пожеж є здатність моделей машинного навчання:

- автоматично виявляти нелінійні та мультиплікативні взаємодії між предикторами без необхідності їх явного попереднього означення дослідником;

- коректно функціонувати в умовах мультиколінеарності ознак, або взаємної корельованості метеорологічних показників, що є типовою властивістю кліматичних даних;

- ефективно обробляти неповні дані з пропусками, що є неминучим наслідком атмосферних перешкод та технічних збоїв, розглянутих у підрозділі 1.2.2;

- масштабуватися на великі обсяги різномірних геопросторових спостережень без втрати обчислювальної ефективності.

Окремим і надзвичайно важливим аспектом наукової новизни розробленої системи є свідомо інтеграція принципів пояснювального штучного інтелекту (Explainable Artificial Intelligence, XAI) як рівноправного архітектурного компонента поряд із власне прогностичною

моделлю. У сфері екологічного моніторингу та управління надзвичайними ситуаціями ця властивість набуває особливого практичного значення з кількох причин.

По-перше, відповідальні управлінські рішення – щодо превентивної евакуації населення, перерозподілу обмежених ресурсів пожежогасіння між територіями або введення особливого протипожежного режиму – вимагають не лише точного числового прогнозу рівня ризику, а й зрозумілого фахівцю пояснення того, які саме умови середовища спричинили цей прогноз. Система, що видає виключно абстрактне число «висока ймовірність пожежі» без розкриття причинно-наслідкового механізму (чи це спричинено аномально низькою вологістю, тривалою відсутністю опадів, критичним станом рослинності, чи поєднанням кількох факторів одночасно), має суттєво нижчу практичну цінність для особи, яка приймає рішення, порівняно із системою, що чітко ідентифікує та ранжує домінуючі фактори ризику.

По-друге, методи пояснювального ШІ слугують потужними інструментами наукової верифікації адекватності побудованої моделі. Якщо аналіз внеску окремих ознак виявляє, що модель приймає рішення переважно на основі предикторів, позбавлених фізичного чи екологічного змісту, це є прямим сигналом про можливу методологічну проблему – витік інформації з цільової змінної в навчальні ознаки, помилки розмітки даних або некоректну специфікацію моделі. Натомість узгодженість виявлених моделлю закономірностей із усталеними фізичними уявленнями про природу пожежного ризику (розглянутими в підрозділі 1.1.1) є вагомим аргументом на користь коректності та надійності побудованої системи.

По-третє, у контексті можливого практичного впровадження подібних систем органами державного управління чи екологічними службами, вимога інтерпретованості алгоритмічних рішень дедалі частіше набуває не лише методологічного, а й нормативного характеру, узгоджуючись із

загальносвітовою тенденцією до підвищення прозорості автоматизованих систем підтримки прийняття рішень.

1.5 Об'єкти дослідження: регіони Марокко та Алжиру як репрезентативні території Північної Африки

1.5.1 Обґрунтування вибору географічного регіону дослідження

Вибір Королівства Марокко та Алжирської Народної Демократичної Республіки як основних географічних об'єктів дослідження є методологічно обґрунтованим рішенням, що спирається на кілька вагомих наукових аргументів.

По-перше, як зазначено в підрозділі 1.1.1, регіон Північної Африки належить до зони підвищеної та зростаючої пожежної небезпеки внаслідок унікального поєднання середземноморського клімату, близькості пустелі Сахара та специфічних орографічних умов гірської системи Атлас. Це робить регіон надзвичайно репрезентативним для дослідження екстремальних проявів пожежного ризику – на відміну від помірних кліматичних зон, де пожежна небезпека має менш виражений сезонний та екстремальний характер.

По-друге, регіон характеризується наявністю задокументованих, відкритих масивів історичних даних про пожежні події, що поєднують супутникові спостереження та наземні метеорологічні вимірювання – необхідна передумова для застосування методів машинного навчання з учителем (supervised learning), що вимагають значних обсягів розмічених історичних спостережень.

По-третє, дослідження одночасно двох географічно суміжних, але не ідентичних країн – Марокко (з виходом до Атлантичного океану і Середземного моря) та Алжиру (переважно середземноморське узбережжя з глибокою континентальною, передсахарською частиною) – створює

унікальну можливість порівняльного аналізу: наскільки добре закономірності, виявлені моделлю машинного навчання для одного регіону, узагальнюються на сусідній регіон зі схожим, але не тотожним кліматичним профілем. Таке порівняння є важливим внеском у розуміння меж застосовності (генералізації) моделей, навчених на географічно-специфічних даних, що детально розглядається в розділах 3 та 4 даної роботи.

1.5.2 Фізико-географічна характеристика Марокко як основного об'єкта моделювання і Алжиру як додаткового об'єкта валідації

Королівство Марокко, що виступає основним об'єктом для побудови центральної прогностичної моделі даної роботи, розташоване на крайньому північному заході Африканського континенту і характеризується надзвичайно контрастним рельєфом – від низовинних атлантичних рівнин до високогірних масивів Високого Атласу з вершинами понад 4000 метрів над рівнем моря. Така топографічна різноманітність зумовлює формування кількох відмінних кліматичних мікрзон у межах однієї країни: вологий середземноморський клімат на півночі, посушливий континентальний клімат у центральних регіонах та аридний, передсахарський клімат на південному сході.

Лісовий покрив Марокко зосереджений переважно в гірських масивах Середнього та Високого Атласу, а також у прибережних зонах. Домінуючими видами є коркові дуби, атлаські кедри та хвойні породи, що, як зазначалося в підрозділі 1.1.1, мають підвищену природну займистість через високий вміст ефірних олій. Саме поєднання значної топографічної та кліматичної варіативності в межах відносно компактної території робить Марокко надзвичайно цінним полігоном для навчання моделі машинного навчання, здатної вловлювати залежність пожежного ризику від географічного положення спостереження – властивість, що детально

аналізується в розділі 4 на основі результатів SHAP-аналізу побудованої моделі.

Алжирська Народна Демократична Республіка, найбільша за площею країна Африканського континенту, є другим географічним об'єктом дослідження, що використовується для перевірки можливості розширення розробленого підходу на суміжний регіон. На відміну від Марокко, північна, середземноморська частина Алжиру (зокрема регіони Беджаїя та Сіді-Бель-Аббес, дані щодо яких використано в даній роботі) характеризується дещо іншим профілем рослинності та мікрокліматичних умов, хоча й належить до того самого загального середземноморського кліматичного поясу.

Використання даних з історичного, добре задокументованого набору спостережень пожежної небезпеки для цих алжирських провінцій, що традиційно застосовується науковою спільнотою для верифікації методів на основі Канадської системи індексів FWI (розглянутої в підрозділі 1.1.2), дозволяє в межах даної роботи реалізувати додатковий, незалежний контур моделювання, що спирається на класичні метеорологічні показники пожежної небезпеки, та порівняти ефективність такого підходу з основною моделлю, побудованою на значно ширшому наборі супутникових та геопросторових предикторів для Марокко. Детальний опис обох наборів даних та порівняння побудованих моделей наведено в розділах 3 та 4.

1.6 Постановка задачі дослідження

Метою кваліфікаційної роботи є розробка інтелектуальної інформаційної системи прогнозування ризику виникнення лісових пожеж на основі методів машинного навчання, що інтегрує супутникові та метеорологічні дані, забезпечує високу точність та просторову деталізацію прогнозу, а також надає користувачу прозоре пояснення факторів, що вплинули на отриманий результат.

Для досягнення поставленої мети в роботі вирішуються такі задачі:

- систематизація теоретичних засад предметної галузі – факторів виникнення та поширення лісових пожеж, специфіки досліджуваного географічного регіону, еволюції методів прогнозування пожежної небезпеки;

- критичний аналіз існуючих інформаційних систем моніторингу лісових пожеж та виявлення притаманних їм функціональних і технологічних обмежень;

- обґрунтування вибору та підготовка масивів вхідних даних для побудови моделей машинного навчання для двох географічних регіонів – Марокко та Алжиру;

- розробка, навчання та порівняльне оцінювання кількох алгоритмів машинного навчання для задачі класифікації пожежного ризику;

- інтеграція методів пояснювального штучного інтелекту для забезпечення інтерпретованості отриманих моделлю прогнозів;

- проєктування та реалізація програмного забезпечення – інтерактивної інформаційної системи, що надає кінцевому користувачу доступ до прогностичних та аналітичних можливостей розробленої моделі.

Об’єктом дослідження є процеси виникнення та поширення лісових пожеж як динамічні екологічні явища, що формуються під впливом сукупності метеорологічних, топографічних та геопросторових факторів у регіонах Північної Африки.

Предметом дослідження є методи, моделі та інформаційні технології машинного навчання, призначені для аналізу геопросторових і метеорологічних даних та прогнозування на їх основі рівня пожежної небезпеки, а також методи забезпечення інтерпретованості результатів такого прогнозування.

Практична цінність отриманих результатів полягає у створенні повнофункціонального інтелектуального інструментарію Forest Fire Intelligence System (FFIS). На відміну від існуючих аналогів, розроблена система поєднує високу точність ансамблевого моделювання з технологіями

пояснювального штучного інтелекту, що дозволяє не лише констатувати факт наявності пожежного ризику, але й детально інтерпретувати причинно-наслідкові зв'язки екологічних факторів для кожного конкретного регіону.

Наукова новизна отриманих результатів полягає в удосконаленні методів екологічного моніторингу шляхом інтеграції ансамблевих алгоритмів машинного навчання з технологіями пояснювального ШІ. Вперше для специфічних кліматичних умов Північної Африки запропоновано підхід, який не просто генерує ймовірнісний прогноз пожежі, а створює динамічний профіль ризику з детальним кількісним обґрунтуванням впливу кожного локального метеорологічного та просторового фактора.

Впровадження розробленої системи має вагомий соціально-економічний ефект. Перехід від реактивної моделі пожежогасіння до проактивного превентивного моніторингу дозволить оптимізувати розподіл ресурсів рятувальних служб, мінімізувати інфраструктурні збитки та зменшити обсяги викидів парникових газів. Крім того, гнучка модульна архітектура забезпечує легку масштабованість системи: вона дозволяє без суттєвих структурних змін інтегрувати нові джерела супутникових даних та адаптувати навчені моделі для інших пожежонебезпечних регіонів світу.

Сформульовані мета, задачі, об'єкт та предмет дослідження визначають вектор подальшого викладу матеріалу роботи. Для успішної побудови та практичної реалізації заявленої інтелектуальної системи виникає необхідність глибокого дослідження математичного апарату та специфіки сучасних підходів до аналізу даних, що буде зроблено в наступному розділі.

2 ТЕОРЕТИЧНІ ЗАСАДИ МАШИННОГО НАВЧАННЯ ТА ПОЯСНЮВАЛЬНОГО ШТУЧНОГО ІНТЕЛЕКТУ

2.1 Огляд алгоритмів машинного навчання

У першому розділі роботи було встановлено, що класичні емпіричні підходи до оцінювання пожежної небезпеки не здатні адекватно врахувати нелінійний, багатофакторний характер взаємодії метеорологічних, топографічних та геопросторових предикторів, що зумовлює необхідність застосування методів машинного навчання. Даний розділ присвячено систематичному викладу теоретичних засад, на яких ґрунтується розроблена інтелектуальна система: математичних принципів роботи алгоритмів класифікації, що розглядаються як кандидати для розв'язання поставленої задачі, теоретичних основ методів пояснювального штучного інтелекту, фізичних принципів розрахунку супутникових вегетаційних індексів, а також огляду сучасного технологічного інструментарію для реалізації подібних інтелектуальних систем. Викладений у цьому розділі теоретичний апарат становить методологічну основу для розробки конкретної архітектури та практичної реалізації системи, опис яких наведено в розділах 3 та 4.

2.1.1 Алгоритми бінарної класифікації

Завдання прогнозування ризику виникнення лісових пожеж у межах даного дослідження формалізується як класична задача бінарної класифікації. Її математична суть полягає у віднесенні кожного геопросторового спостереження, представленого вектором метеорологічних та супутникових ознак, до одного з двох взаємовиключних класів: «пожежа» (позитивний клас) або «відсутність пожежі» (негативний клас).

Вибір оптимального алгоритму для розв'язання цієї задачі потребує врахування складної нелінійної природи екологічних даних. У зв'язку з цим, для проведення порівняльного аналізу та побудови інтелектуальної системи було відібрано чотири репрезентативні методи машинного навчання, що еволюціонують від базових статистичних підходів до передових ансамблевих архітектур.

Логістична регресія є одним із найстаріших та найбільш фундаментальних алгоритмів бінарної класифікації [4], що широко застосовується як референтна (baseline) модель при порівняльному аналізі складніших алгоритмів. Незважаючи на свою відносну простоту, логістична регресія залишається важливим інструментом завдяки високій інтерпретованості та обчислювальній ефективності.

В основі методу лежить ідея побудови лінійної комбінації вхідних ознак, яка потім перетворюється у ймовірність належності до класу за допомогою сигмоїдної функції. Сигмоїдна функція має характерну S-подібну форму та відображає будь-яке дійсне число у діапазоні від 0 до 1, що дозволяє інтерпретувати вихід моделі безпосередньо як ймовірність належності спостереження до класу «пожежа» (рисунк 2.1).

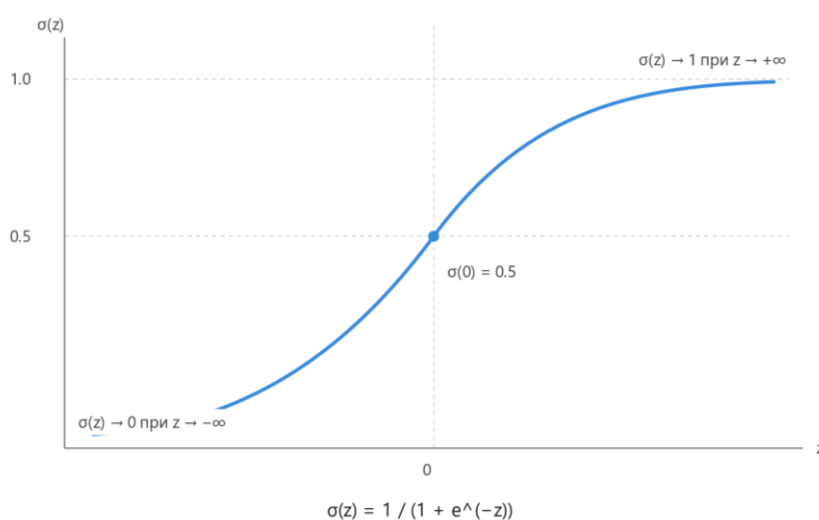


Рисунок 2.1 – Графік сигмоїдної функції та її властивість обмеження значень у діапазоні (0,1)

Навчання коефіцієнтів моделі здійснюється методом максимізації функції правдоподібності, що на практиці еквівалентно мінімізації функції втрат логарифмічної похибки (log loss, бінарної крос-ентропії). Для запобігання перенавчанню типово застосовується регуляризація – додатковий штрафний доданок до функції втрат, що обмежує надмірне зростання коефіцієнтів моделі та підвищує її здатність до узагальнення на нових даних [4].

Принциповим теоретичним обмеженням логістичної регресії є базове припущення про лінійну роздільність класів у просторі перетворених (логіт) ознак [20]. У задачах прогнозування пожежного ризику взаємозв'язки між метеорологічними та геопросторовими предикторами мають яскраво виражений нелінійний характер, що обговорювалося у підрозділі 1.2.3 – наприклад, вплив температури на ризик пожежі суттєво посилюється лише після перевищення певного порогового значення, а ефект сильного вітру критично залежить від одночасного рівня вологості (мультиплікативна взаємодія факторів). Логістична регресія не здатна автоматично виявляти такі взаємодії без явного попереднього конструювання відповідних додаткових ознак дослідником, що зумовлює її використання переважно як референтної моделі.

Таким чином, попри обмежену виразність щодо нелінійних взаємодій, логістична регресія залишається невід'ємним компонентом порівняльного дослідження як еталон простоти та інтерпретованості, відносно якого оцінюється приріст якості більш складних моделей.

Метод опорних векторів або SVM (рисунок 2.2) є потужним алгоритмом класифікації, що базується на пошуку оптимальної розділової гіперплощини в просторі ознак [3]. На відміну від логістичної регресії, яка моделює безпосередньо ймовірність належності до класу, SVM формулює задачу класифікації як задачу геометричної оптимізації – пошуку гіперплощини, що максимізує відстань (margin, розділову смугу) до найближчих представників кожного класу.

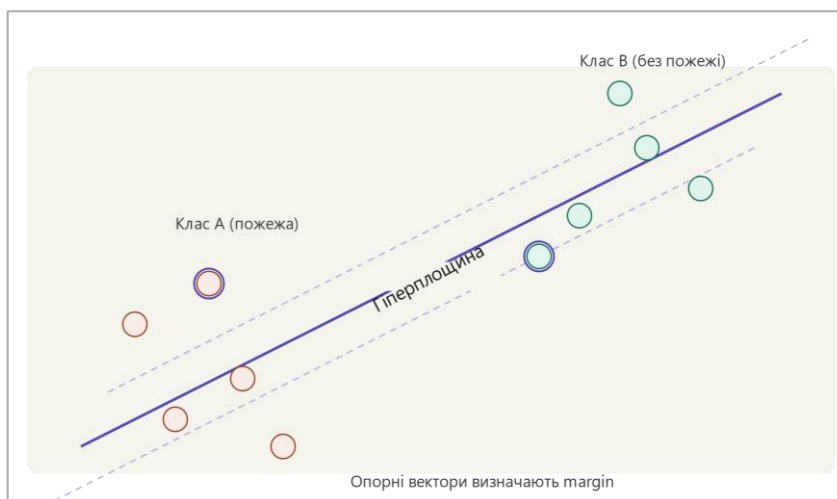


Рисунок 2.2 – Принцип побудови оптимальної розділової гіперплощини та ядрового переходу SVM

Для лінійно нероздільних даних, що є типовою ситуацією для метеорологічних та супутникових спостережень, застосовується так званий «ядровий перехід» (kernel trick) – математичний прийом, що дозволяє неявно відобразити вхідні дані у простір значно вищої розмірності, де лінійне розділення стає можливим, без необхідності явного обчислення координат у цьому просторі. Найбільш широко використовуваною ядровою функцією для задач із складною нелінійною структурою є радіальна базисна функція (Radial Basis Function, RBF), параметр якої визначає «ширину впливу» кожного опорного вектора – великі значення цього параметра призводять до більш локальних, складних розділових поверхонь з підвищеним ризиком перенавчання, тоді як малі значення забезпечують гладкіші, більш узагальнюючі розділові поверхні.

Ключовою практичною проблемою застосування SVM до великих масивів даних є висока обчислювальна складність навчання, що зростає квадратично або навіть кубічно відносно кількості навчальних спостережень через необхідність обробки повної матриці попарних відстаней між усіма точками навчальної вибірки.

Алгоритм Random Forest (рисунок 2.3) належить до класу ансамблевих методів машинного навчання, що базуються на агрегації прогнозів багатьох базових моделей для отримання результату, точнішого та стійкішого, ніж може забезпечити будь-яке окреме дерево [2].

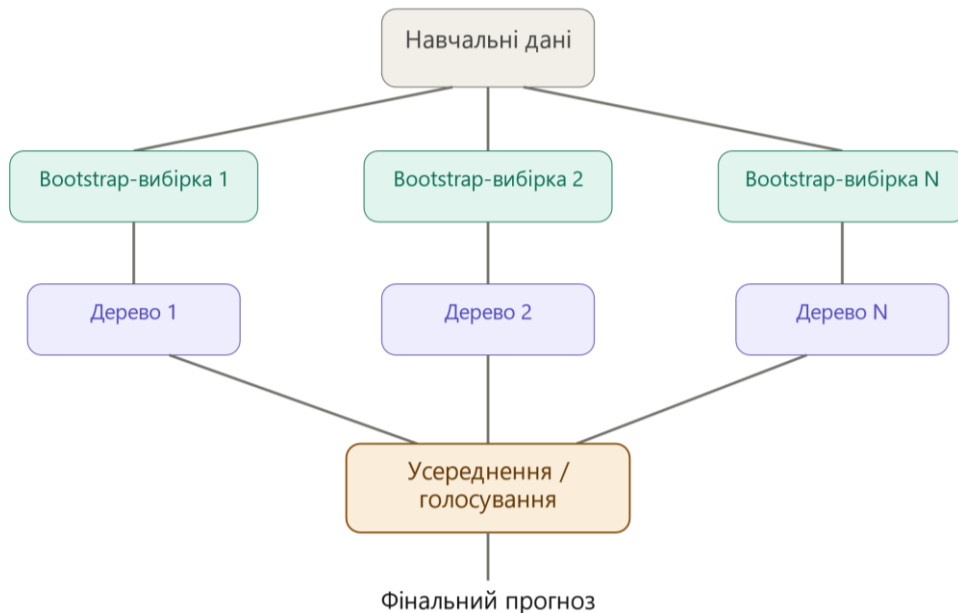


Рисунок 2.3 – Схема ансамблю Random Forest: бутстреп-вибірки та агрегація прогнозів дерев

Дерево рішень будує ієрархічну деревоподібну структуру правил, що рекурсивно поділяють простір ознак на області з максимально однорідним розподілом цільового класу. Кожен поділ обирається за критерієм мінімізації статистичної неоднорідності (наприклад, індексу Джині) у дочірніх вузлах. Окреме дерево рішень, навчене до повного розгалуження, як правило, демонструє надзвичайно високу дисперсію – воно фактично запам'ятовує шум, притаманний конкретній навчальній вибірці, та погано узагальнюється на нові, раніше не бачені спостереження (явище перенавчання).

Random Forest вирішує цю проблему через комбінацію двох незалежних механізмів випадковості. По-перше, метод бутстреп-агрегації

передбачає, що кожне дерево ансамблю навчається на власній випадковій підвибірці навчальних даних, отриманій методом вибору з поверненням, унаслідок чого кожне дерево «бачить» дещо інший підмножину спостережень, що знижує кореляцію між прогнозами окремих дерев. По-друге, при побудові кожного дерева на кожному кроці поділу розглядається не повний набір доступних ознак, а лише випадково обрана їх підмножина, що додатково урізноманітнює структуру дерев ансамблю [2].

Фінальний прогноз моделі формується через усереднення (для задачі оцінювання ймовірності) або мажоритарне голосування (для задачі класифікації) прогнозів усіх дерев ансамблю [2]. Завдяки усередненню значна частина випадкового шуму, притаманного окремим деревам, взаємно компенсується, що призводить до суттєвого зниження загальної дисперсії прогнозу ціною незначного зростання систематичного зміщення (bias-variance trade-off). Саме ця властивість робить Random Forest особливо стійким до перенавчання та до наявності аномальних значень (викидів) у вхідних метеорологічних даних.

Додатковою практичною перевагою Random Forest є вбудована можливість оцінювання важливості ознак без потреби у зовнішніх методах інтерпретації, що реалізується через обчислення середнього зниження індексу Джині при поділах за кожною ознакою, усередненого по всіх деревах ансамблю.

Алгоритм XGBoost є удосконаленою реалізацією методу градієнтного бустингу [1] – техніки ансамблевого навчання, принципово відмінної від Random Forest за способом побудови ансамблю. На відміну від паралельного, незалежного навчання дерев у методі Bagging, у градієнтному бустингу дерева будуються послідовно (ітеративно), причому кожне нове дерево спеціально навчається для виправлення помилок, накопичених попередніми деревами ансамблю.

Загальна модель XGBoost є сумою прогнозів усіх послідовно побудованих дерев (рисунок 2.4). На кожному кроці ітеративного навчання

нове дерево додається до ансамблю таким чином, щоб мінімізувати регуляризовану цільову функцію – суму функції втрат (що враховує наявний прогноз ансамблю на попередньому кроці) та штрафного члена регуляризації, що обмежує надмірну складність нового дерева (кількість листків та величину прогнозованих ними значень).

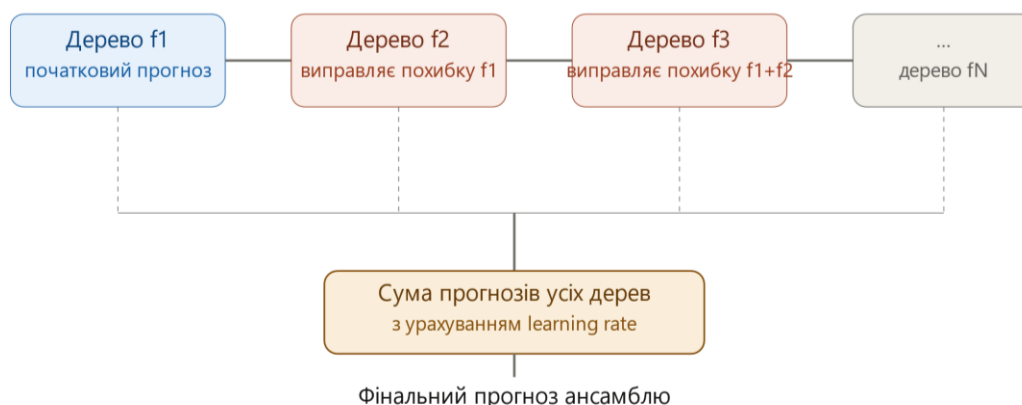


Рисунок 2.4 – Схема послідовної побудови ансамблю дерев у градієнтному бустингу

Ключовою інженерною та математичною інновацією XGBoost є застосування розкладу другого порядку (наближення Тейлора) для апроксимації функції втрат при побудові кожного нового дерева – на відміну від класичного градієнтного бустингу, що використовує лише інформацію про першу похідну (градієнт), XGBoost додатково враховує другу похідну (гесіан) функції втрат, що забезпечує суттєво швидшу та точнішу збіжність оптимізаційного процесу [1].

Серед ключових гіперпараметрів алгоритму, що визначають поведінку моделі, особливе практичне значення мають: швидкість навчання, що масштабує внесок кожного нового дерева в загальний прогноз ансамблю та контролює компроміс між швидкістю збіжності й стійкістю до перенавчання; частки спостережень та ознак, що випадково обираються для побудови кожного окремого дерева (subsampling), які вносять додатковий

елемент стохастичності за аналогією з методом Random Forest; та максимальна глибина окремого дерева, що обмежує складність базових моделей ансамблю.

Поєднання послідовного адаптивного виправлення помилок, використання інформації другого порядку про функцію втрат та комплексної регуляризації робить XGBoost алгоритмом, що на практиці систематично демонструє одну з найвищих точностей серед класичних методів машинного навчання на табличних (структурованих) даних середнього масштабу [1].

2.1.2 Теоретичні основи оцінювання якості класифікаційних моделей

Коректне оцінювання якості моделі бінарної класифікації вимагає застосування комплексу взаємодоповнюючих метрик, оскільки жодна окрема метрика не здатна повністю охарактеризувати якість прогнозування в умовах, типових для задачі пожежного ризику – зокрема, в умовах можливого дисбалансу класів, теоретично розглянутого в підрозділі 1.2.1.

Базовою одиницею оцінювання є матриця помилок – таблиця, що зіставляє прогнозовані моделлю класи з фактичними, виділяючи чотири категорії результатів: істинно позитивні (модель коректно прогнозує пожежу), істинно негативні (модель коректно прогнозує відсутність пожежі), хибно позитивні (модель помилково прогнозує пожежу там, де її немає) та хибно негативні (модель помилково не виявляє реальну пожежу). На основі цих чотирьох величин обчислюється весь подальший комплекс метрик:

– точність (Accuracy) визначається як частка правильних прогнозів від загальної кількості спостережень. Незважаючи на інтуїтивну зрозумілість, ця метрика має істотне теоретичне обмеження, докладно розглянуте в підрозділі 1.2.1: за умов значного дисбалансу класів висока точність може досягатися моделлю, що практично ігнорує клас меншості, що робить

Ассурасу ненадійним самостійним індикатором якості для задач подібного типу;

- точність позитивного класу (Precision) показує частку справді позитивних випадків серед усіх спостережень, які модель класифікувала як позитивні, тобто характеризує, наскільки можна довіряти позитивному прогнозу моделі;

- повнота (Recall, або чутливість) показує частку реальних позитивних випадків, які моделі вдалося правильно виявити, тобто характеризує здатність моделі не пропускати справжні випадки пожежі. Для задачі превентивного моніторингу пожежного ризику метрика Recall має особливе практичне значення, оскільки хибно негативний прогноз, або пропущена існуюча пожежна небезпека) є значно небезпечнішою помилкою, ніж хибно позитивний (зайва увага до безпечної території);

- оскільки між Precision та Recall існує природний компроміс, а саме підвищення порогу класифікації типово збільшує Precision ціною зниження Recall, і навпаки, для отримання єдиної збалансованої метрики застосовується F1-міра – гармонічне середнє точності та повноти, що набуває високого значення лише за умови одночасно високих значень обох складових метрик;

- площа під ROC-кривою (ROC-AUC) є інтегральною метрикою, що оцінює здатність моделі коректно ранжувати позитивні спостереження вище за негативні для всього можливого діапазону порогових значень класифікації одночасно, а не лише для одного фіксованого порогу. Теоретичною перевагою цієї метрики є її незалежність від конкретного обраного порогу прийняття рішення та від абсолютного співвідношення класів у вибірці, що робить її особливо придатною для порівняння моделей між собою;

- логарифмічна втрата (Log Loss) є метрикою, що оцінює якість не дискретних класифікаційних рішень моделі, а безпосередньо прогнозованих ймовірностей, суттєво штрафуючи модель за впевнені, але неправильні

прогнози сильніше, ніж за обережні помилки, що робить цю метрику теоретично спорідненою з функцією втрат, на основі якої здійснюється навчання моделей логістичної регресії та градієнтного бустингу, розглянутих у підрозділі 2.1.1.

2.2 Теоретичні основи методів пояснювального штучного інтелекту

2.2.1 Проблема «чорного ящика» та метод SHAP

Сучасні ансамблеві моделі машинного навчання, розглянуті в підрозділі 2.1, здатні досягати високої точності прогнозування саме завдяки своїй здатності виявляти складні нелінійні взаємодії між десятками й сотнями вхідних ознак. Однак ця складність має зворотний бік – внутрішня логіка прийняття рішень подібною моделлю стає практично непрозорою для людини-аналітика. Модель, що складається з сотень окремих дерев рішень, кожне з яких містить десятки умовних розгалужень за багатьма ознаками, не піддається безпосередньому «читанню» та інтерпретації людиною – звідси й виникла усталена в науковій літературі метафора «чорного ящика».

Проблема непрозорості моделей має кілька важливих практичних наслідків для предметної галузі прогнозування пожежного ризику, детально обговореної в розділі 1. По-перше, відсутність пояснення ускладнює формування довіри до системи з боку фахівців-практиків, які звикли працювати з прозорими, фізично інтерпретованими індексами на кшталт класичного FWI, де кожен компонент має чіткий і зрозумілий фізичний зміст. По-друге, неможливість пояснення суттєво ускладнює діагностику помилкових прогнозів моделі – без спеціалізованих інструментів інтерпретації неможливо встановити, чи є конкретна помилка випадковою статистичною похибкою, чи симптомом системної методологічної проблеми (наприклад, витоку інформації з цільової змінної, некоректно сформованої ознаки чи артефакту навчальної вибірки). Саме ці міркування

зумовлюють необхідність інтеграції в розроблювану систему спеціалізованого модуля пояснювального штучного інтелекту.

Метод SHAP, запропонований у 2017 році, є одним із найбільш теоретично обґрунтованих та широко застосовуваних методів пояснення прогнозів моделей машинного навчання [5]. Концептуальною основою методу є теорія кооперативних ігор, а саме поняття значення Шеплі (Shapley value), запропоноване Ллойдом Шеплі ще у середині XX століття для розв'язання задачі справедливого розподілу спільного виграшу між учасниками кооперативної гри.

В аналогії з кооперативною грою, кожна вхідна ознака моделі, наприклад, температура, вегетаційний індекс або швидкість вітру, розглядається як окремий «гравець», а прогноз моделі для конкретного спостереження інтерпретується як спільний «виграш», що має бути справедливо розподілений між цими гравцями-ознаками пропорційно до їхнього реального внеску у формування результату (рисунок 2.5). Значення Шеплі для кожної ознаки обчислюється як середній маргінальний внесок цієї ознаки, усереднений за всіма можливими порядками послідовного «додавання» ознак до моделі та зважений відповідно до кількості можливих комбінацій (коаліцій) такого розміру.

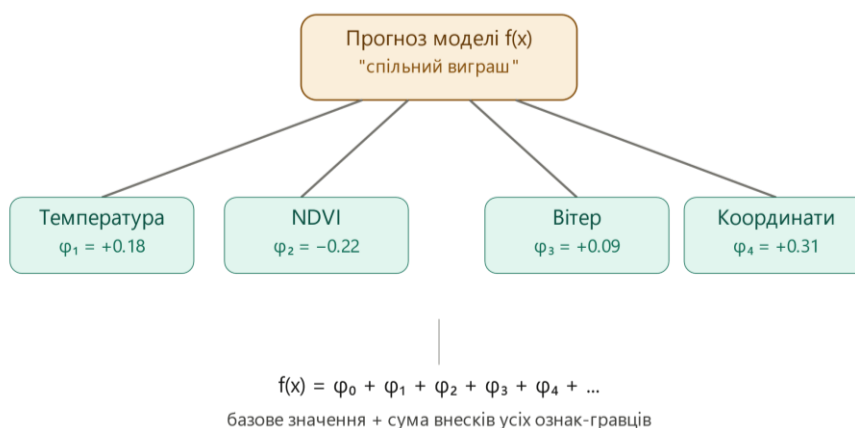


Рисунок 2.5 – Ілюстрація розподілу спільного «виграшу» між гравцями-ознаками за теорією кооперативних ігор

Інтуїтивно подібний підхід дозволяє коректно врахувати взаємодії між ознаками: внесок ознаки «швидкість вітру» у підвищення прогнозованого ризику може бути суттєво більшим за наявності одночасно низької вологості, ніж за її відсутності, і метод значень Шеплі коректно відображає цю залежність завдяки усередненню по всіх можливих коаліціях ознак, а не лише по їхньому ізольованому впливу.

Метод SHAP базується на кількох фундаментальних аксіомах, що гарантують теоретичну коректність та математичну унікальність отриманого розподілу внесків ознак [5]:

- аксіома локальної точності (Local Accuracy) стверджує, що сума значень SHAP усіх ознак для конкретного спостереження, додана до базового (середнього очікуваного) прогнозу моделі, точно дорівнює фактичному прогнозу моделі для цього спостереження. Ця властивість має важливе практичне значення: вона гарантує, що сума всіх «факторів впливу», відображених в інтерфейсі системи користувачу, точно відтворює фінальну прогнозовану ймовірність, не залишаючи нічого «незрозумілого» поза межами пояснення;

- аксіома відсутності впливу (Missingness) встановлює, що якщо певна ознака відсутня у вхідних даних або не справляє жодного впливу на прогноз для конкретного спостереження, її значення SHAP дорівнює нулю – інтуїтивно очевидна, але формально необхідна властивість коректного методу пояснення;

- аксіома узгодженості (Consistency) гарантує, що якщо дві ознаки роблять однаковий або більший маргінальний внесок у прогноз для всіх можливих коаліцій порівняно з іншою моделлю, їхнє значення SHAP не може зменшитися. Ця властивість забезпечує об'єктивну порівнюваність відносної важливості ознак як у межах однієї моделі, так і між різними моделями.

2.2.2 Алгоритм TreeSHAP для дерево-базованих моделей

Пряме обчислення точних значень Шеплі за загальною формулою теорії кооперативних ігор має експоненційну обчислювальну складність відносно кількості вхідних ознак, що робить його практично нездійсненним для моделей із десятками чи сотнями предикторів, типовими для задач, подібних до розглянутої в даній роботі [5].

Для подолання цього обмеження спеціально для дерево-базованих моделей, зокрема, Random Forest та XGBoost, розглянутих у підрозділі 2.1.1, розроблено спеціалізований алгоритм TreeSHAP, що досягає поліноміальної обчислювальної складності за рахунок використання внутрішньої структури самого дерева рішень для точного, а не наближеного обчислення необхідних умовних очікувань, без потреби явного перебору експоненційної кількості можливих коаліцій ознак.

Це досягається завдяки тому, що при «виключенні» певної ознаки з умовного очікування алгоритм ефективно усереднює прогнози по всіх шляхах дерева, що не залежать від цієї ознаки, використовуючи статистику навчальних спостережень, накопичену в кожному листку дерева під час навчання.

Практична реалізація методу SHAP передбачає два взаємодоповнюючих рівні аналізу. Глобальне пояснення (Global Explanation) полягає в обчисленні середнього абсолютного значення SHAP для кожної ознаки, усередненого по репрезентативній вибірці спостережень, що дозволяє побудувати загальний рейтинг важливості всіх ознак моделі та визначити, які фактори в принципі найбільше впливають на прогнози моделі.

Локальне пояснення (Local Explanation), своєю чергою, надає повний вектор значень SHAP для одного конкретного спостереження, що дозволяє відповісти на питання, чому модель видала саме такий прогноз для цієї конкретної комбінації вхідних умов – типово подібне пояснення

візуалізується у вигляді так званої waterfall-діаграми (рисунок 2.6), що демонструє послідовний внесок кожного фактора у формування фінального прогнозу відносно базового (середнього) значення.

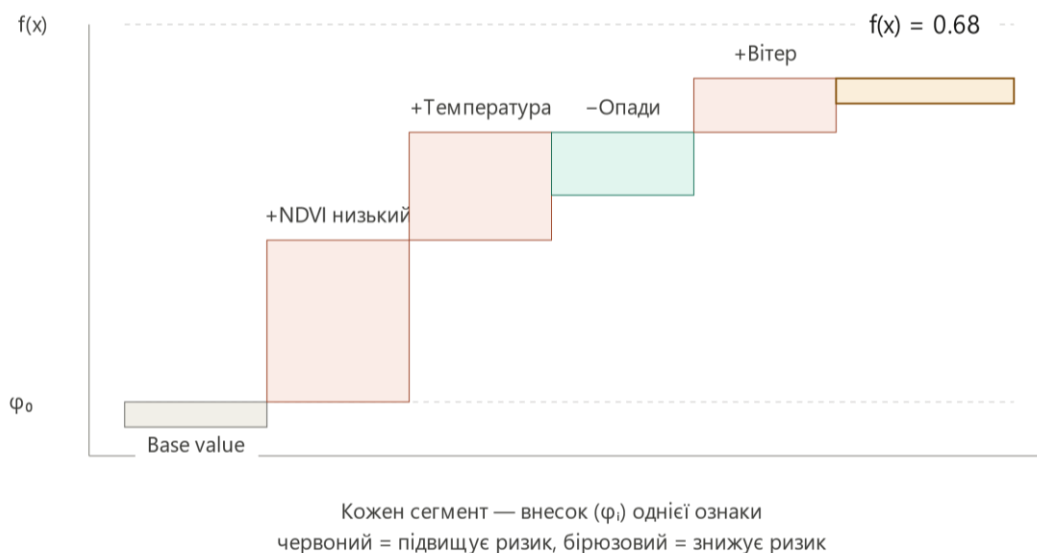


Рисунок 2.6 – Приклад waterfall-діаграми як форми локального пояснення SHAP

2.3 Теоретичні основи аналізу геопросторових та супутникових даних

2.3.1 Принципи розрахунку спектральних вегетаційних індексів

Як зазначалося в підрозділі 1.1.3, спектральні вегетаційні індекси є похідними показниками, що обчислюються на основі комбінації відбиття електромагнітного випромінювання в різних спектральних діапазонах, зафіксованого мультиспектральними супутниковими сенсорами. Теоретичною основою таких індексів є відмінність спектральних властивостей здорової фотосинтезуючої рослинності порівняно з пригніченою, сухою рослинністю або оголеним ґрунтом (рисунок 2.7).

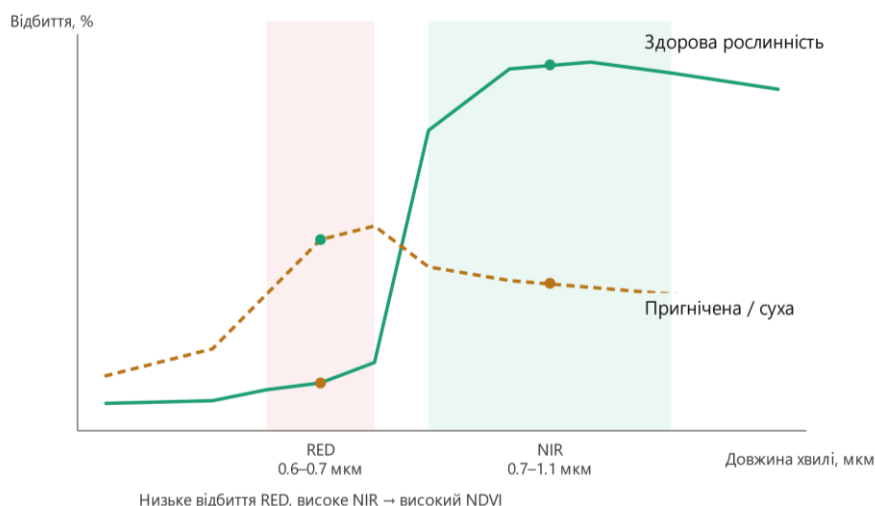


Рисунок 2.7 – Спектральні криві відбиття здорової та пригніченої рослинності в червоному та ближньому ІЧ-діапазонах

Нормалізований різницевий вегетаційний індекс (NDVI) обчислюється як нормалізована різниця між відбиттям у ближньому інфрачервоному діапазоні та відбиттям у видимому червоному діапазоні спектра, віднесена до їхньої суми [6]. Значення індексу теоретично варіюються в діапазоні від мінус одиниці до плюс одиниці: значення, близькі до нуля або від’ємні, типово відповідають водним поверхням, оголеному ґрунту чи штучним покриттям; помірні позитивні значення відповідають розрідженій або пригніченій рослинності; високі значення, які наближені до 0,7–0,9, характерні для густої, фізіологічно здорової рослинності з активним фотосинтезом.

Принципова цінність NDVI для задачі прогнозування пожежного ризику полягає в тому, що індекс опосередковано, але надійно відображає водний статус рослинного покриву – зниження вмісту вологи в листі та стеблах рослин внаслідок посухи чи теплового стресу призводить до зміни їхніх оптичних властивостей і, відповідно, до вимірюваного зниження значення NDVI задовго до того, як цей стрес стане візуально помітним для людини-спостерігача. Це робить динаміку NDVI у часі (а не лише його

миттєве значення) цінним випереджальним індикатором накопиченого водного дефіциту рослинності.

Спорідненим, проте теоретично відмінним показником є нормалізований різницевий водний індекс (NDWI) [6], що використовує комбінацію ближнього інфрачервоного та короткохвильового інфрачервоного діапазонів спектра, чутливих безпосередньо до вмісту рідкої води в тканинах листя, а не лише до структурних і пігментних характеристик рослинності, як у випадку класичного NDVI. Поєднання обох індексів дозволяє отримати повніше уявлення про фізіологічний стан рослинного покриву порівняно з використанням будь-якого з них окремо.

Окрім NDVI, сучасна практика дистанційного зондування передбачає використання додаткових спектральних індексів, що характеризують різні аспекти стану рослинного покриву. Зокрема, індекс EVI (Enhanced Vegetation Index) є менш чутливим до атмосферних аерозолів та насичення сигналу в зонах густої рослинності, що підвищує точність оцінки в умовах задимлення від наближених пожеж.

2.3.2 Теоретичні засади часових лагових ознак та кумулятивних метеорологічних показників

Як зазначалося в підрозділі 1.2.4, миттєвий («поточний») стан метеорологічних умов є недостатнім предиктором пожежного ризику, оскільки фізичні процеси накопичення та втрати води рослинністю і ґрунтом мають виражений кумулятивний, інерційний характер. Теоретичне обґрунтування використання так званих лагових ознак (lag-features) – значень метеорологічних показників, виміряних не в поточний момент часу, а з певним часовим зсувом (наприклад, один, три чи сім днів тому) – спирається на принципи моделювання часових рядів та фізику вологообміну в природних екосистемах.

Кожен шар горючого матеріалу (тонка лісова підстилка, товстіший шар нерозкладеної органіки, деревина) характеризується власною інерційністю (часовою сталою) реакції на зміну атмосферних умов: тонкі горючі матеріали втрачають або накопичують вологу протягом годин, тоді як товстіші шари органіки реагують на зміну умов протягом днів і навіть тижнів. Це теоретичне положення лежить в основі багатокомпонентної структури класичних індексів пожежної небезпеки (зокрема, системи FWI, розглянутої в підрозділі 1.1.2), де окремі компоненти (FFMC, DMC, DC) моделюють вологість шарів горючого матеріалу з різною часовою інерційністю за допомогою системи диференціальних рівнянь.

У контексті методів машинного навчання аналогічний ефект досягається не через явне моделювання фізичних диференціальних рівнянь вологообміну, а через надання моделі безпосереднього доступу до історичних значень метеорологічних показників за кілька попередніх днів у формі окремих вхідних ознак, а також агрегованих статистик (ковзних середніх) за різні часові вікна – тижневі, місячні, сезонні. Це дозволяє моделі самостійно «навчитися» оптимальній ваговій функції залежності поточного ризику від історії метеорологічних умов, без необхідності апріорного припущення конкретної математичної форми цієї залежності, що є додатковою ілюстрацією переваги data-driven підходу, обговореної в підрозділі 1.4.1.

2.3.3 Геопросторова інтерполяція та просторова прив'язка даних

Як зазначалося в підрозділі 1.2.4, узгодження різнорідних за просторовою структурою джерел даних – точкових наземних метеорологічних вимірювань та неперервних растрових супутникових продуктів – є самостійним методологічним викликом. Теоретичною основою вирішення цієї проблеми є методи геопросторової інтерполяції, що дозволяють оцінити значення показника в довільній точці простору на

основі відомих значень у скінченній множині інших точок (метеостанцій або центрів растрових комірок).

Найпростіший теоретичний підхід полягає у використанні географічних координат (широти та довготи) точки спостереження безпосередньо як вхідних ознак моделі машинного навчання. На перший погляд це може здаватися надлишковим, проте в умовах територіально неоднорідного клімату (типового для гірських та прибережних регіонів, розглянутих у підрозділі 1.5) географічні координати опосередковано кодують інформацію про належність точки до певної кліматичної мікрозони – фактично виконуючи роль проксі-змінної (*proxy variable*) для довгострокових, усталених кліматичних характеристик території, які не повністю відображаються короткостроковими метеорологічними вимірюваннями. Теоретичне обґрунтування та емпіричне підтвердження значущості такого підходу детально розглядаються за результатами аналізу важливості ознак побудованої моделі в розділі 4.

2.4 Огляд сучасного технологічного інструментарію для розробки інтелектуальних систем аналізу даних

Мова програмування Python є де-факто галузевим стандартом для розробки систем машинного навчання та аналізу даних, що зумовлено кількома факторами. По-перше, екосистема Python пропонує один із найбільш зрілих та повних наборів спеціалізованих бібліотек для наукових обчислень, статистичного аналізу та машинного навчання серед усіх мов програмування, що дозволяє реалізувати повний цикл розробки інтелектуальної системи – від завантаження та попередньої обробки даних до навчання моделей та їх інтерпретації – в єдиному, узгодженому програмному середовищі.

По-друге, синтаксична лаконічність та читабельність мови Python суттєво знижують поріг входження та пришвидшують ітеративний процес

дослідницької розробки, що є особливо важливим на етапі експериментального підбору архітектури моделі та гіперпараметрів. По-третє, Python забезпечує однаково потужну підтримку як обчислювально-інтенсивних задач (завдяки інтеграції з оптимізованими бібліотеками, написаними на низькорівневих мовах), так і задач розробки користувацьких інтерфейсів та вебзастосунків, що дозволяє реалізувати завершену інтелектуальну систему без необхідності міжмовної інтеграції різних технологічних стеків.

Сучасна екосистема Python для машинного навчання включає кілька категорій спеціалізованих інструментів. Бібліотеки для маніпуляції структурованими табличними даними забезпечують ефективні, векторизовані операції над великими обсягами спостережень, включаючи фільтрацію, агрегацію, об'єднання різних джерел даних та обчислення похідних статистичних показників – функціональність, що є основою для реалізації методів попередньої обробки даних, теоретично розглянутих у підрозділах 1.2 та 2.3.

Бібліотеки для класичного машинного навчання надають уніфікований програмний інтерфейс для реалізації широкого спектра алгоритмів – від лінійних моделей та методу опорних векторів до ансамблевих методів, розглянутих у підрозділі 2.1, – а також інструментарій для коректного розбиття даних на навчальну й тестову вибірки, стандартизації ознак та обчислення метрик якості моделей. Окремі спеціалізовані бібліотеки реалізують конкретні високопродуктивні алгоритми (зокрема, оптимізовану реалізацію градієнтного бустингу XGBoost), а також методи пояснювального штучного інтелекту, теоретично розглянуті в підрозділі 2.2.

Для відображення результатів прогнозування на інтерактивній географічній карті в сучасній практиці розробки геоінформаційних застосунків широко застосовуються спеціалізовані картографічні бібліотеки, що надають програмний доступ до функціональності побудови

інтерактивних карт – масштабування, нанесення маркерів, побудови теплових карт густини розподілу досліджуваного показника та накладання додаткових тематичних шарів, включно з безпосередньою інтеграцією супутникових тайлових сервісів. Такий підхід дозволяє поєднати в єдиному інтерфейсі результати прогнозування моделі машинного навчання з безпосередньою візуалізацією супутникових даних, що використовувалися для побудови цього прогнозу, забезпечуючи додаткову прозорість та довіру до системи з боку кінцевого користувача.

Для реалізації користувацького інтерфейсу сучасних аналітичних систем у галузі науки про дані дедалі ширшого застосування набувають спеціалізовані фреймворки, що дозволяють розробникам створювати повноцінні інтерактивні вебзастосунки засобами тієї самої мови програмування, що використовується для реалізації аналітичного ядра системи, без необхідності окремої розробки клієнтської частини засобами традиційних вебтехнологій. Подібний підхід суттєво скорочує час розробки прототипів інтелектуальних систем та забезпечує безпосередню, безшовну інтеграцію навченої моделі машинного навчання з інтерактивним інтерфейсом користувача, що є особливо цінним при розробці систем дослідницького та демонстраційного характеру, подібних до розробленої в межах даної роботи.

2.5 Порівняльний теоретичний аналіз розглянутих підходів

Узагальнення теоретичних властивостей чотирьох розглянутих у підрозділі 2.1 алгоритмів класифікації дозволяє сформулювати цілісне уявлення про їхні порівняльні переваги та обмеження в контексті задачі прогнозування пожежного ризику (таблиця 2.1), що є необхідною методологічною передумовою для обґрунтованого вибору конкретної моделі або моделей на етапі практичної реалізації, викладеної в розділах 3 та 4.

Таблиця 2.1 – Порівняльний теоретичний аналіз алгоритмів класифікації за ключовими властивостями

Критерій	Logistic Regression	SVM (RBF)	Random Forest	XGBoost
Тип моделі	Лінійна	Ядрова	Ансамбль (bagging)	Ансамбль (boosting)
Здатність до нелінійних залежностей	Низька	Висока	Висока	Дуже висока
Стійкість до мультиколінеарності ознак	Низька	Середня	Висока	Висока
Обчислювальна складність навчання	Низька	Висока (квадратична / кубічна)	Помірна	Помірна
Потреба у попередній стандартизації ознак	Так	Так	Ні	Ні
Інтерпретованість «з коробки»	Висока	Низька	Середня	Середня (потребує ХАІ)
Стійкість до аномальних значень (викидів)	Низька	Середня	Висока	Висока
Механізм побудови ансамблю	–	–	Паралельний (bagging)	Послідовний (boosting)

Як випливає з теоретичного розгляду, лінійні моделі (логістична регресія) забезпечують найвищу інтерпретованість і обчислювальну ефективність ціною обмеженої здатності моделювати нелінійні взаємодії; ядрові методи (SVM) долають це обмеження за рахунок суттєвого зростання обчислювальної складності; ансамблеві методи на основі дерев рішень, а саме Random Forest і XGBoost, теоретично поєднують здатність до моделювання складних нелінійних залежностей з прийнятною обчислювальною ефективністю, проте потребують додаткової інтеграції методів пояснювального штучного інтелекту, теоретично розглянутих у підрозділі 2.2, для компенсації притаманної їм проблеми «чорного ящика».

Аналогічним чином, теоретичний розгляд методів пояснювального ШІ (підрозділ 2.2) та геопросторового аналізу даних (підрозділ 2.3)

демонструє їхню взаємодоповнюваність: якщо методи на кшталт SHAR забезпечують інтерпретованість на рівні окремих ознак моделі, то усвідомлене теоретичне розуміння фізичної природи самих ознак, наприклад, вегетаційних індексів, лагових метеорологічних показників, геопросторових координат, є необхідною передумовою для змістовної інтерпретації результатів такого аналізу, що детально продемонстровано на основі результатів практичної реалізації моделі в розділі 4.

Сформований у другому розділі теоретичний базис об'єднує три фундаментальні компоненти майбутньої системи: методи нелінійного моделювання структурованих даних, математичний апарат пояснювального штучного інтелекту та фізико-екологічну специфіку просторово-часових предикторів. Синергія цих підходів дозволяє не лише нівелювати ризики перенавчання моделей на складних геопросторових вибірках, але й трансформувати високовірогідні ансамблеві рішення з категорії концептуального «чорного ящика» у прозорий, інтерпретований інструмент підтримки прийняття управлінських рішень.

Визначені переваги, обмеження та компроміси розглянутих підходів закладають необхідну методологічну основу для наступного етапу дослідження. У третьому розділі буде представлено формальну математичну специфікацію задачі прогнозування пожежного ризику, деталізовано цільові функції оптимізації обраних алгоритмів, а також розроблено архітектурні UML-моделі та схеми інформаційних потоків інтелектуальної системи.

3 МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА МЕТОДОЛОГІЯ ПРОГНОЗУВАННЯ РИЗИКУ ПОЖЕЖІ

У розділах 1 та 2 роботи було обґрунтовано актуальність задачі, проаналізовано предметну галузь та викладено теоретичні засади алгоритмів машинного навчання й методів пояснювального штучного інтелекту, що розглядаються як кандидати для розв'язання поставленої задачі. Даний розділ присвячено формальній математичній постановці задачі прогнозування пожежного ризику, точному математичному опису обраних моделей та метрик їх оцінювання, а також детальній методології підготовки вхідних даних і формальній архітектурі розробленої системи, що в сукупності становить безпосередню методологічну основу для практичної реалізації, описаної в розділі 4.

3.1 Формальна постановка задачі прогнозування

Задача прогнозування ризику виникнення лісової пожежі формалізується як задача бінарної класифікації з учителем (supervised binary classification). Нехай задано навчальну вибірку:

$$D = \{(x_i, y_i)\}_{i=1}^N, \text{ де } x_i \in R^n, \quad (3.1)$$

де $x_i \in R^n$ – вектор вхідних ознак i -го спостереження;

$y_i \in \{0, 1\}$ – відповідна йому цільова мітка (1 – зафіксовано пожежу, 0 – пожежу не зафіксовано);

N – загальна кількість спостережень у вибірці;

n – розмірність простору ознак.

Метою навчання є побудова функції (моделі) $f: R^n \rightarrow [0,1]$, що апроксимує умовну ймовірність належності спостереження до класу «пожежа» за заданим вектором ознак:

$$f(x) \approx P(y = 1 | x). \quad (3.2)$$

Якість наближення функції f оцінюється шляхом мінімізації обраної функції втрат L на навчальній вибірці:

$$f^* = \arg \min_f 1/N \sum_{i=1}^N L(y_i, f(x_i)), \quad (3.3)$$

де f^* – оптимальна (навчена) модель, а конкретний вигляд функції втрат L залежить від обраного алгоритму (детально розглянуто в підрозділі 3.2 для кожної з чотирьох моделей).

Для прийняття дискретного класифікаційного рішення (бінарної мітки) на основі неперервної прогнозованої ймовірності застосовується пороговий оператор:

$$\hat{y} = 1, \text{ якщо } f(x) \geq t; \hat{y} = 0, \text{ якщо } f(x) < t, \quad (3.4)$$

де $t \in [0,1]$ – поріг класифікації (типове значення $t = 0.5$).

У межах даної роботи задача додатково розширюється до багаторівневої класифікації ризику шляхом розбиття неперервного інтервалу прогнозованої ймовірності на чотири категоріальні рівні:

$$\text{Рівень}(x) = \begin{cases} \text{LOW, якщо } f(x) < 0.25 \\ \text{MEDIUM, якщо } 0.25 \leq f(x) < 0.55. \\ \text{HIGH, якщо } 0.55 \leq f(x) < 0.75 \end{cases} \quad (3.5)$$

Вектор вхідних ознак x для основної моделі (регіон Марокко) складається з $n = 158$ компонент, що належать до шести функціональних груп (таблиця 3.1), формально означених як:

$$x = (x_{\text{geo}}, x_{\text{sat}}, x_{\text{temp}}, x_{\text{meteo}}, x_{\text{roll}}, x_{\text{eng}})^T, \quad (3.6)$$

де $x_{\text{geo}} \in R^3$ – географічні ознаки (широта, довгота, відстань до моря);

$x_{\text{sat}} \in R^2$ – супутникові індекси (NDVI, вологість ґрунту);

$x_{\text{temp}} \in R^4$ – календарні ознаки;

$x_{\text{meteo}} \in R^{77}$ – лагові метеорологічні ознаки (11 змінних $\times$ 7 днів);

$x_{\text{roll}} \in R^{49}$ – ковзні агреговані статистики (7 змінних $\times$ 7 часових вікон);

$x_{\text{eng}} \in R^5$ – синтетичні інженерні ознаки.

Для допоміжної моделі (регіон Алжир) використовується суттєво менший за розмірністю вектор ознак $x' \in R^{10}$, що складається з чотирьох безпосередньо вимірюваних метеорологічних показників та шести компонентів Канадської системи індексів пожежної небезпеки FWI, формальний опис яких наведено в підрозділі 3.4.

Таблиця 3.1 – Структура вхідного простору ознак для моделей Марокко та Алжиру

Група ознак	Марокко (n=158)	Алжир (n=10)
Географічні	3 (широта, довгота, відстань до моря)	–
Супутникові	2 (NDVI, вологість ґрунту)	–
Часові (календарні)	4 (день тижня, день року, вихідний, свято)	–
Метеорологічні лагові	77 (11 змінних $\times$ 7 днів)	4 (температура, вологість, вітер, опади)
Ковзні агреговані статистики	49 (7 змінних $\times$ 7 вікон)	–
Інженерні (синтетичні)	5	–
Компоненти FWI	–	6 (FFMC, DMC, DC, ISI, BUI, FWI)
Разом	158	10

3.2 Математичний апарат алгоритмів класифікації

3.2.1 Лінійні та ядрові моделі

Модель логістичної регресії визначає прогнозовану ймовірність як композицію лінійного перетворення вхідних ознак та сигмоїдної функції активації [4]:

$$f(x) = \sigma(z) = \frac{1}{1+e^{(-z)}}, \text{ де } z = \beta_0 + \sum_{i=1}^n \beta_i x_i, \quad (3.7)$$

де β_0 – вільний член (зсув);

β_i – коефіцієнт регресії для i -ї ознаки.

Навчання коефіцієнтів β здійснюється шляхом мінімізації функції втрат логарифмічної похибки (бінарної крос-ентропії) з доданим членом L2-регуляризації:

$$L(\beta) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(f(x_i)) + (1 - y_i) \log(1 - f(x_i))] + \frac{1}{(2C)} \sum_{j=1}^n \beta_j^2, \quad (3.8)$$

де C – обернений до сили регуляризації гіперпараметр (менші значення C відповідають сильнішій регуляризації).

Насамперед навчання SVM зводиться до задачі квадратичного програмування – пошуку гіперплощини, що максимізує розділову смугу (margin) між класами за наявності штрафних змінних нечіткості для лінійно нероздільних спостережень. Нелінійне розділення забезпечується ядровою функцією RBF, що неявно визначає скалярний добуток у просторі вищої розмірності [3]:

$$K(x_i, x_j) = \phi(x_i)^T \cdot \phi(x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad (3.9)$$

де $\gamma > 0$ – параметр ширини ядра.

Фінальне рішення приймається на основі суми зважених ядрових відстаней до опорних векторів (рисунок 3.1). Детальний вивід двоїстої задачі оптимізації виходить за межі даної роботи й наведений у класичних працях з теорії SVM [3].

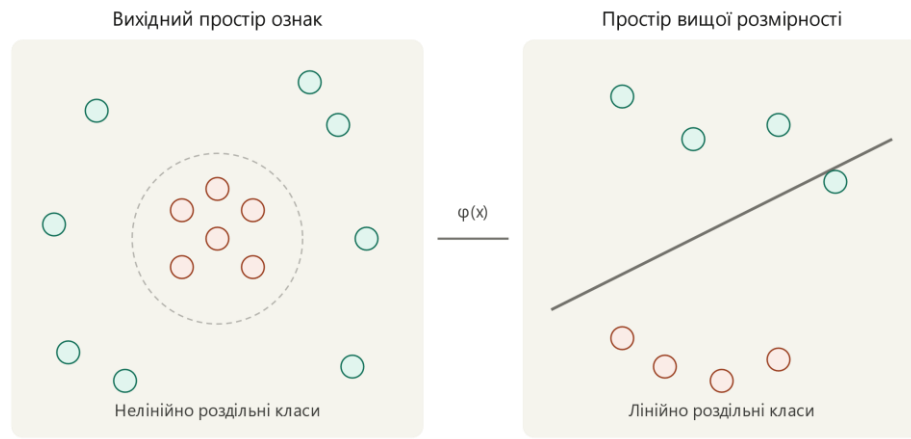


Рисунок 3.1 – Принцип розділення класів у вихідному просторі ознак (ліворуч) та у трансформованому просторі вищої розмірності після застосування ядрового переходу (праворуч)

3.2.2 Ансамблеві моделі

Прогноз моделі Random Forest для задачі оцінювання ймовірності формується як усереднення прогнозів T незалежно навчених дерев рішень [2]:

$$f(x) = \frac{1}{T} \sum_{t=1}^T h_t(x), \quad (3.10)$$

де $h_t(x)$ – прогноз t -го дерева, навченого на бутстреп-вибірці D_t , отриманій випадковим вибором N спостережень з поверненням з вихідної навчальної вибірки D .

Кожен поділ вузла дерева обирається шляхом мінімізації індексу Джині для двох дочірніх піднабірок:

$$G(S) = 1 - \sum_{k=0}^1 p_k^2, \quad (3.11)$$

де p_k – частка спостережень класу k у вузлі S .

$$(j^*, s^*) = \arg \min_{(j,s)} \left[\frac{|S_L|}{|S|} \cdot G(S_L) + \frac{|S_R|}{|S|} \cdot G(S_R) \right], \quad (3.12)$$

де j – індекс ознаки, за якою здійснюється поділ;

s – порогове значення поділу;

S_L та S_R – ліва й права дочірні підмножини.

Модель XGBoost формується як адитивний ансамбль K послідовно побудованих дерев [1]:

$$\hat{y}_i^{(K)} = \sum_{k=1}^K f_k(x_i), \quad f_k \in F, \quad (3.13)$$

де F – простір регресійних дерев (CART).

На кожній ітерації t нове дерево f_t додається до ансамблю шляхом мінімізації регуляризованої цільової функції (рисунки 3.2):

$$L^{(t)} = \sum_{i=1}^N l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \quad (3.14)$$

$$\Omega(f_t) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^T w_j^2, \quad (3.15)$$

де T – кількість листків дерева f ;

w_j – вага j -го листка;

γ та λ – параметри регуляризації.

Застосовуючи розклад Тейлора другого порядку до функції втрат відносно прогнозу попередньої ітерації, цільова функція апроксимується як:

$$L^{(t)} \approx \sum_{i=1}^N \left[g_i f_t(x_i) + \frac{h_i f_t(x_i)^2}{2} \right] + \Omega(f_t), \quad (3.16)$$

де $g_i = \left(\partial \ell(y_i, \hat{y}_i^{(t-1)}) \right) / \left(\partial \hat{y}_i^{(t-1)} \right)$ – градієнт (перша похідна);

$h_i = \left(\partial^2 \ell(y_i, \hat{y}_i^{(t-1)}) \right) / \left(\partial (\hat{y}_i^{(t-1)})^2 \right)$ – гесіан (друга похідна) функції

втрат для i -го спостереження.

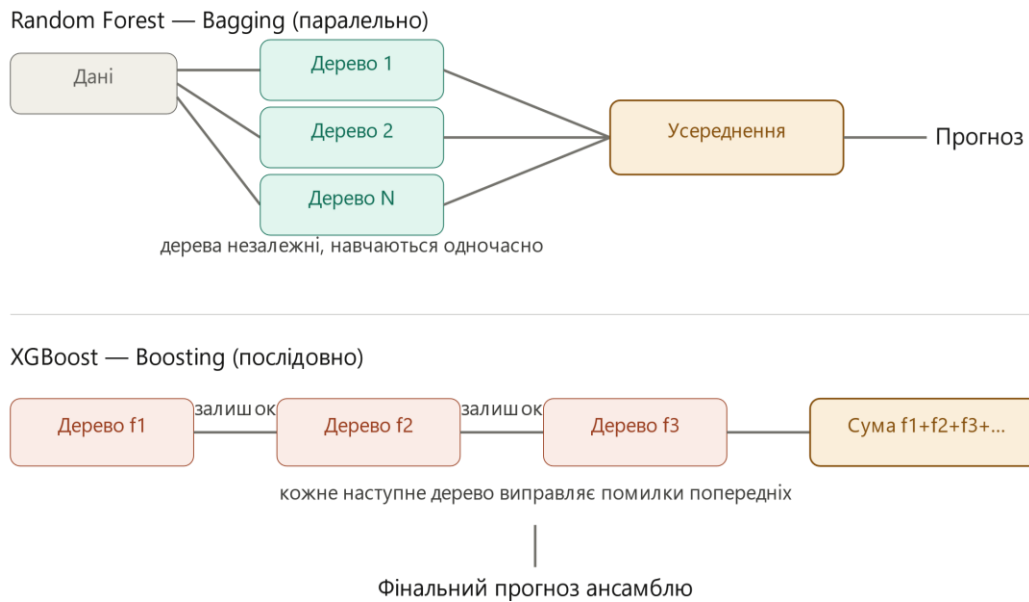


Рисунок 3.2 – Порівняльна схема побудови ансамблю: паралельний бегінг (Random Forest) проти послідовного бустингу (XGBoost)

Для наглядності було сформовано таблицю порівняння гіперпараметрів моделей класифікації (таблиця 3.2).

Таблиця 3.2 – Порівняння ключових гіперпараметрів чотирьох моделей класифікації

Модель	Ключові гіперпараметри
Logistic Regression	$C = 1.0$, $\text{max_iter} = 1000$, $\text{solver} = \text{lbfgs}$
SVM (RBF)	$\text{kernel} = \text{rbf}$, $C = 1.0$, $\text{gamma} = \text{scale}$
Random Forest	$n_estimators = 300$, $\text{max_depth} = 12$, $\text{min_samples_leaf} = 10$, $\text{max_features} = \text{sqrt}$
XGBoost	$n_estimators = 500$, $\text{max_depth} = 6$, $\text{learning_rate} = 0.05$, $\text{subsample} = 0.8$, $\text{colsample_bytree} = 0.8$

3.3 Математичний апарат методу SHAP

Значення Шеплі для i -ї ознаки визначається як середньозважений маргінальний внесок цієї ознаки за всіма можливими коаліціями ознак, які її не містять [5]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{(|S|!(n-|S|-1)!)}{n!} [f(S \cup \{i\}) - f(S)], \quad (3.17)$$

де $f(S)$ – очікуваний прогноз моделі за умови відомих значень лише ознак з коаліції S .

Метод однозначно задовольняє аксіому локальної точності (адитивності), що гарантує точне розкладання прогнозу моделі на суму внесків усіх ознак та базового значення:

$$f(x) = \phi_0 + \sum_{i=1}^n \phi_i, \text{ де } \phi_0 = E[f(x)]. \quad (3.18)$$

Пряме обчислення формули 3.17 має експоненційну складність $O(2^n)$; алгоритм TreeSHAP, застосований у даній роботі, досягає поліноміальної складності $O(T \cdot L \cdot D^2)$ (T – кількість дерев, L – кількість листків, D – глибина дерева) за рахунок аналітичного обчислення умовних очікувань безпосередньо зі структури дерева. Глобальна важливість ознаки обчислюється усередненням абсолютних SHAP-значень по вибірці (рисунок 3.3):

$$I_i = \frac{1}{M} \sum_{j=1}^M |\phi_i^{(j)}|. \quad (3.19)$$

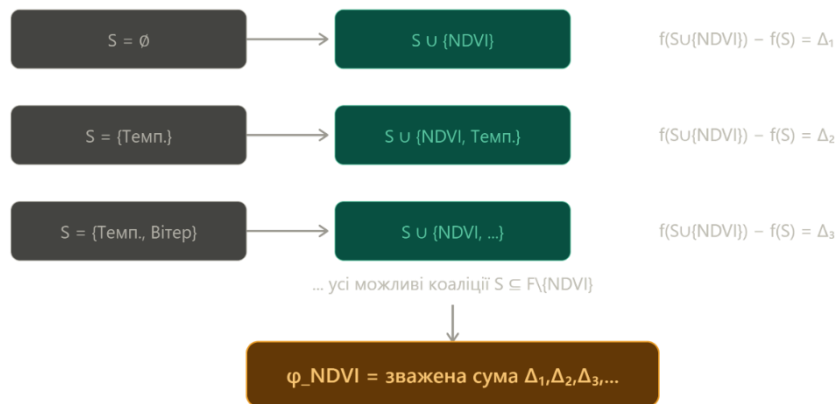


Рисунок 3.3 – Схема обчислення значень Шеплі: розподіл прогнозу моделі на внески окремих ознак-гравців

3.4 Формули супутникових індексів та інженерних ознак

$$NDVI = \frac{(\rho_{(NIR)} - \rho_{(RED)})}{(\rho_{(NIR)} + \rho_{(RED)})}, \quad (3.20)$$

$$NDWI = \frac{(\rho_{(NIR)} - \rho_{(SWIR)})}{(\rho_{(NIR)} + \rho_{(SWIR)})}, \quad (3.21)$$

Створені синтетичні ознаки моделі Марокко можна побачити на таблиці 3.3.

Таблиця 3.3 – П'ять синтетичних (інженерних) ознак моделі Марокко: формули та інтерпретація

Ознака	Формула	Інтерпретація
Температурний діапазон	$\Delta T = T_{(\max, \text{lag1})} - T_{(\min, \text{lag1})}$	Добова амплітуда, ознака сухої континентальної погоди
Взаємодія температури і вітру	$I = T_{(\text{avg}, \text{lag1})} \times V_{(\text{witer}, \text{lag1})}$	Мультиплікативний ефект двох факторів
Індекс посушливості	$D = \frac{P_{(\text{lag1})}}{(T_{(\text{avg}, \text{lag1})} + 1)}$	Дефіцит вологи відносно температури
Прапор аномалії NDVI	$\text{flag} = 1,$ якщо $NDVI < Q_{0.15}$	Індикатор критично пригніченої рослинності
Комплексний індекс ризику	$S = 0.4T - 0.3P + 0.2G - 0.1 \cdot NDVI$	Зважена апріорна оцінка ризику

Шість компонентів Канадської системи FWI (FFMC, DMC, DC, ISI, BUI, FWI), що використовуються як готові вхідні ознаки для допоміжної моделі (Алжир), обчислюються за стандартизованою методикою Канадської лісової служби на основі температури, вологості, вітру та опадів і в межах даної роботи виступають вхідними даними датасету, а не предметом самостійної реалізації; ієрархію їх взаємного обчислення наведено в таблиці 3.4.

Таблиця 3.4 – Ієрархія компонентів системи FWI та вхідні метеорологічні показники для кожного з них

Компонент	Назва	Вхідні показники	Часова інерційність
FFMC	Fine Fuel Moisture Code	Температура, вологість, вітер, опади	≈ 16 годин
DMC	Duff Moisture Code	Температура, вологість, опади	≈ 12 діб
DC	Drought Code	Температура, опади	≈ 52 доби
ISI	Initial Spread Index	FFMC, швидкість вітру	похідний від FFMC
BUI	Buildup Index	DMC, DC	похідний від DMC, DC
FWI	Fire Weather Index	ISI, BUI	підсумковий індекс

3.5 Формули метрик оцінювання якості моделей

На основі матриці помилок (TP, TN, FP, FN) визначається комплекс метрик якості класифікації:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN), \quad (3.22)$$

$$Precision = TP/(TP + FP), Recall = TP/(TP + FN), \quad (3.23)$$

$$F1 = 2 \cdot (Precision \cdot Recall)/(Precision + Recall), \quad (3.24)$$

$$LogLoss = -\frac{1}{N} \sum_{i=1}^N [y_i \log(f(x_i)) + (1 - y_i) \log(1 - f(x_i))], \quad (3.25)$$

$$AUC = P(f(x^+) > f(x^-)). \quad (3.26)$$

3.6 Формальна архітектура системи

Розроблена інформаційна система побудована за принципом розділення відповідальності (Separation of Concerns) і складається з чотирьох логічно відокремлених рівнів (рисунок 3.4):

- рівня даних (зберігання вихідних датасетів та навчених артефактів моделей);
 - рівня обробки та навчання (завантаження даних, передобробка, навчання й оцінювання моделей);
 - рівня пояснювального аналізу (обчислення та кешування SHAP-значень);
 - рівня користувацького інтерфейсу (інтерактивний вебзастосунок).
- Дані рухаються односпрямовано від рівня обробки до рівня інтерфейсу, що забезпечує незалежність кожного рівня від внутрішньої реалізації сусідніх та дозволяє швидкодію інтерфейсу незалежно від тривалості процесу навчання моделей.

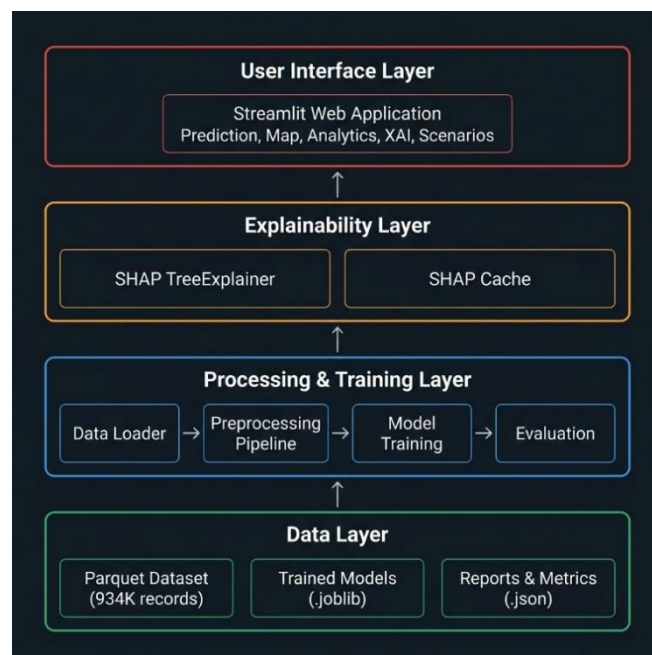


Рисунок 3.4 – Чотирирівнева архітектура системи з потоками даних між рівнями

Такий підхід до побудови архітектури забезпечує високу масштабованість та модульність системи. Наприклад, у разі необхідності оновлення алгоритму машинного навчання (скажімо, переходу від Random Forest до новішої версії XGBoost) або модифікації методів попередньої

обробки даних, не потрібно вносити жодних змін до коду користувацького інтерфейсу. Ізоляція рівня пояснювального аналізу (XAI) дозволяє незалежно оптимізувати та кешувати ресурсоемні обчислення SHAP-значень, що є критично важливим для забезпечення швидкого відгуку вебзастосунку під час інтерактивної взаємодії з користувачем.

Функціональні можливості системи з погляду кінцевого користувача формалізовано у вигляді UML-діаграми (рисунок 3.5) прецедентів використання (Use Case Diagram).

Діаграма охоплює сценарії: виконання прогнозу ризику для довільної точки одного з двох регіонів; перегляд інтерактивної геопросторової карти; проведення розвідувального аналізу даних; порівняння якості моделей; отримання SHAP-пояснення прогнозу; запуск сценарної симуляції умов; перегляд автоматично згенерованого зведеного звіту.

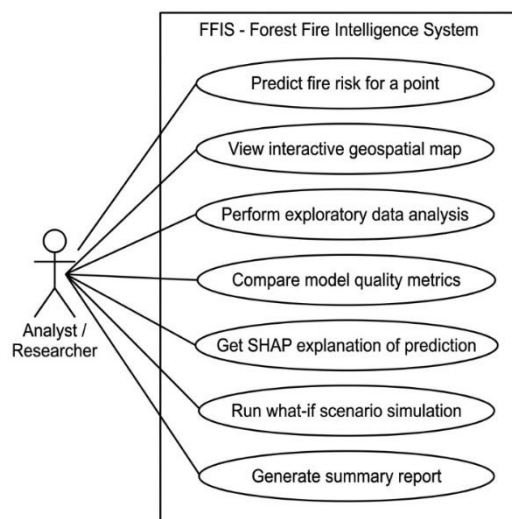


Рисунок 3.5 – Діаграма прецедентів використання (Use Case Diagram) системи

Наведена діаграма прецедентів формалізує всі ключові сценарії взаємодії користувача із системою та слугує основою для визначення функціональних вимог до кожного модуля вебзастосунку, що реалізовані в розділі 4.

Статичну структуру програмних модулів та залежності між ними формалізує діаграма компонентів: компонент користувацького інтерфейсу залежить від компонента управління станом сесії, який, своєю чергою, звертається до компонента доступу до даних (завантажені артефакти моделей і звіти), компонента аналітичного ядра (модулі навчання та інференсу моделей) і компонента пояснювального аналізу (рисунок 3.6).

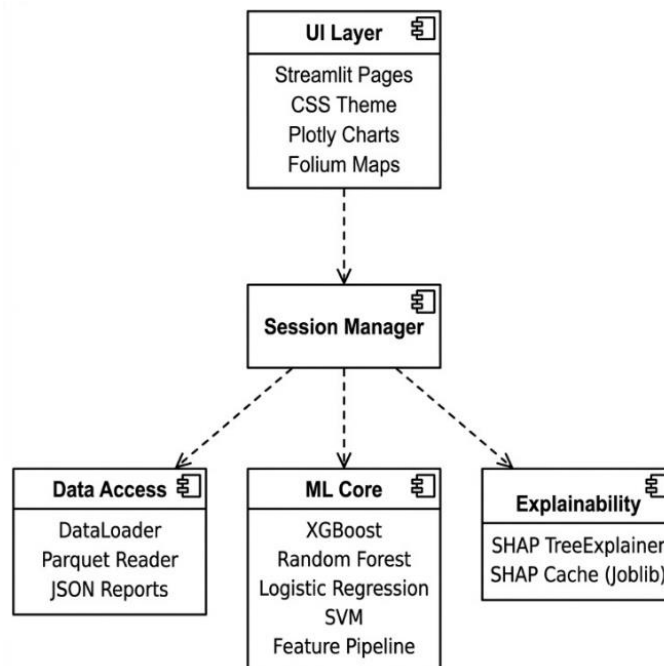


Рисунок 3.6 – Діаграма компонентів (Component Diagram) системи

Центральним зв'язуючим елементом архітектури є компонент управління станом сесії (Session Manager). Оскільки сучасні вебінтерфейси за своєю природою працюють за принципом «запит-відповідь» і часто не зберігають стан між діями користувача, цей менеджер відповідає за безпечне утримання в оперативній пам'яті масивних артефактів моделей (серіалізованих об'єктів XGBoost та SHAP-експлейнерів). Це дозволяє уникнути повторного завантаження гігабайтів даних з жорсткого диска при кожному новому кліку користувача по карті, зводячи час реакції аналітичного ядра до мілісекунд.

Подібна компонентна структура забезпечує можливість незалежного масштабування кожного модуля системи та спрощує інтеграцію нових алгоритмів аналізу або додаткових джерел геопросторових даних без порушення цілісності існуючої функціональності.

Динаміку взаємодії компонентів системи для ключового сценарію – виконання прогнозу для точки, заданої користувачем, – формалізує діаграма послідовності: введення параметрів користувачем → виклик функції прогнозування → завантаження серіалізованої моделі → формування повного вхідного вектора ознак → обчислення прогнозованої ймовірності моделлю → визначення категоріального рівня ризику → відображення результату користувачу (рисунок 3.7).

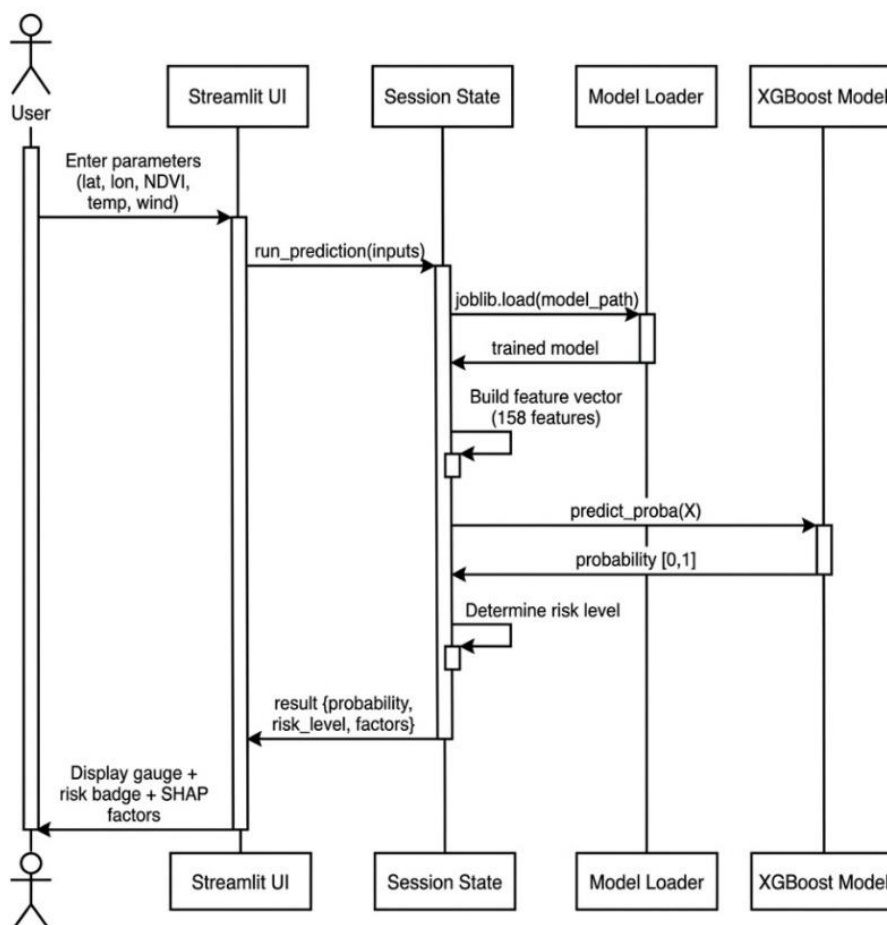


Рисунок 3.7 – Діаграма послідовності (Sequence Diagram) для сценарію прогнозування ризику

Трансформацію даних на шляху від вихідного датасету до фінального прогнозу формалізує діаграма потоків даних (DFD) першого рівня деталізації: вихідний датасет → процес завантаження та валідації → очищений набір даних → процес передобробки та інженерії ознак → матриця ознак → процес навчання моделі → навчена модель і метрики якості → процес обчислення пояснень → глобальні й локальні SHAP-пояснення → відображення в інтерфейсі користувача (рисунок 3.8).

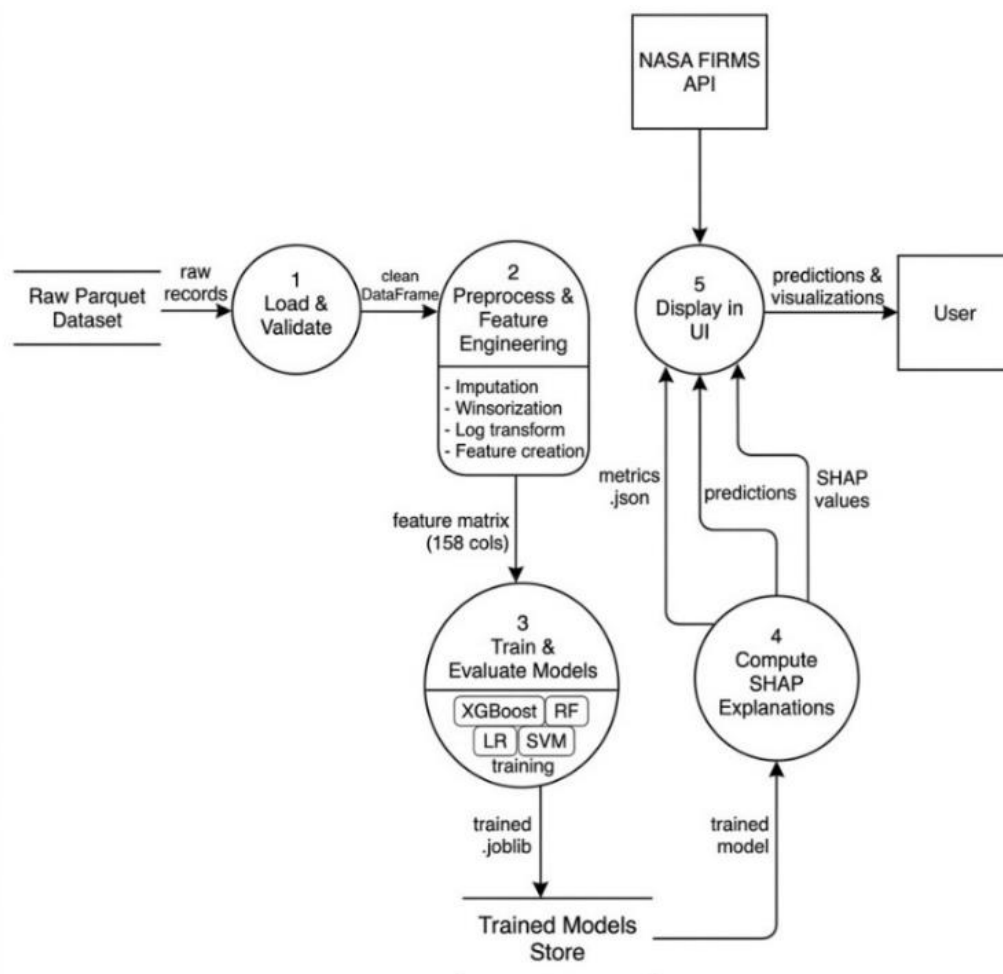


Рисунок 3.8 – Діаграма потоків даних (Data Flow Diagram) системи

Діаграма DFD ілюструє концепцію інформаційного конвеєра (data pipeline), яка є галузевим стандартом у сучасних системах машинного навчання. Вихідні дані проходять строгу послідовність трансформацій, де

найважливішим вузлом є процес інженерії ознак. Саме на цьому етапі розрізнені показники температури, вітру та вегетаційних індексів перетворюються на складні лагові змінні та ковзні агреговані статистики.

Описана строга конвеєризація процесів гарантує повну відтворюваність результатів моделювання та унеможлиблює критичну для машинного навчання проблему витоку даних між навчальною і тестовою вибірками. Крім того, збереження проміжних результатів обробки у вигляді послідовних артефактів дозволяє системі миттєво генерувати високоточні прогнози в режимі реального часу, не перевантажуючи обчислювальні потужності. Таким чином, розроблений комплекс формальних UML-моделей, DFD-схем та архітектурних рішень створює надійний інженерний базис для етапу практичної програмної реалізації інформаційної системи.

4 РЕАЛІЗАЦІЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ ФУНКЦІОНУВАННЯ СИСТЕМИ

У розділах 1–3 роботи обґрунтовано актуальність задачі, викладено теоретичні засади та формалізовано математичний і методологічний апарат розробленої системи. Даний розділ присвячено безпосередньому опису програмної реалізації системи Forest Fire Intelligence System (FFIS), результатам навчання та порівняльного оцінювання моделей, аналізу пояснювального ШІ, опису користувацького інтерфейсу та комплексному тестуванню розробленого програмного продукту.

4.1 Технологічний стек реалізації

Програмна реалізація системи виконана мовою програмування Python 3.11 з використанням таких ключових бібліотек (таблиця 4.1):

- pandas та NumPy – для завантаження й векторизованої обробки табличних даних;
- scikit-learn – для реалізації пайплайну передобробки, моделей Logistic Regression, SVM, Random Forest та обчислення метрик [2];
- XGBoost – для реалізації основної моделі градієнтного бустингу [1];
- SHAP – для обчислення пояснень моделі методом TreeSHAP;
- joblib – для серіалізації навчених моделей;
- Streamlit – для реалізації інтерактивного вебзастосунку;
- Plotly – для побудови інтерактивних графіків;
- Folium та streamlit-folium – для геопросторової візуалізації на інтерактивній карті;
- Ollama – для розгортання локальної великої мовної моделі (LLM) та генерації текстових AI-рекомендацій у вебзастосунку.

Таблиця 4.1 – Технологічний стек системи FFIS за функціональними категоріями

Категорія	Технології
Мова програмування	Python 3.11
Обробка даних	pandas, NumPy
Машинне навчання	scikit-learn, XGBoost
Пояснювальний ШІ	SHAP (TreeExplainer)
Серіалізація моделей	joblib
Вебінтерфейс	Streamlit
Візуалізація даних	Plotly
Геопросторова візуалізація	Folium, streamlit-folium, NASA GIBS
Локальна LLM (рекомендації)	Ollama (llama3.2:3b)

4.2 Структура датасетів та опис вхідних даних

Основним джерелом даних для навчання моделей слугує датасет Morocco Wildfire Prediction (таблиця 4.2) обсягом 934 586 спостережень та 278 вихідних стовпців у бінарному форматі parquet (тип даних float32). Цільова змінна is_fire має ідеально збалансований розподіл (50% / 50%). Для навчання використано стратифіковану вибірку обсягом 200 000 спостережень (по 100 000 кожного класу), що зберігає вихідний баланс і забезпечує статистичну репрезентативність при прийнятному часі навчання.

Для розширення дослідження на суміжний географічний регіон використано датасет Algerian Forest Fires Dataset (таблиця 4.2) [16], що містить 243 коректних спостереження для двох провінцій Алжиру – Bejaia та Sidi-Bel Abbes – за період з червня по вересень 2012 року. Цільова мітка Classes (fire / not fire) перетворена на бінарну змінну is_fire з природним балансом 56% / 44%.

Набір ознак включає чотири безпосередньо вимірних метеопказники (температура, відносна вологість, швидкість вітру, опади) та шість компонентів Канадської системи FWI (FFMC, DMC, DC, ISI, BUI, FWI), теоретично описаних у підрозділі 3.4.3.

Таблиця 4.2 – Порівняльна характеристика двох використаних датасетів

Параметр	Марокко	Алжир
Кількість спостережень	934 586 (вибірка 200 000)	243
Кількість вихідних ознак	278	14
Кількість ознак моделі	158	10
Баланс класів	50% / 50%	56% / 44%
Наявність геокоординат	Так (точкові)	Ні (тільки назва регіону)
Наявність супутникових індексів	Так (NDVI, вологість ґрунту)	Ні
Часовий діапазон	Багаторічний	Один сезон (2012)

У межах вебзастосунку реалізовано інтерактивну сторінку розвідувального аналізу даних (EDA), що включає три типи аналізу: гістограми та боксплоти розподілу ключових ознак (NDVI, вологість ґрунту, температура, опади, вітер, точка роси) з розбиттям за класами цільової змінної; часовий аналіз сезонного розподілу частоти пожеж по місяцях та роках; кореляційну матрицю чотирнадцяти основних числових ознак (рисунки 4.1–4.3).

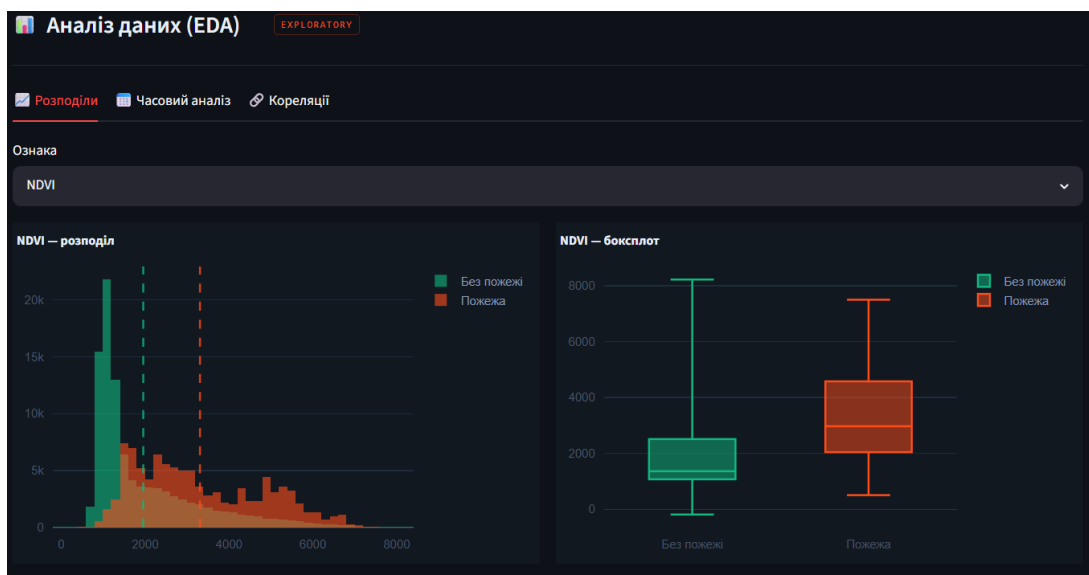


Рисунок 4.1 – Гістограма розподілу NDVI з розбиттям за класами is_fire



Рисунок 4.2 – Сезонний розподіл частоти пожеж по місяцях року

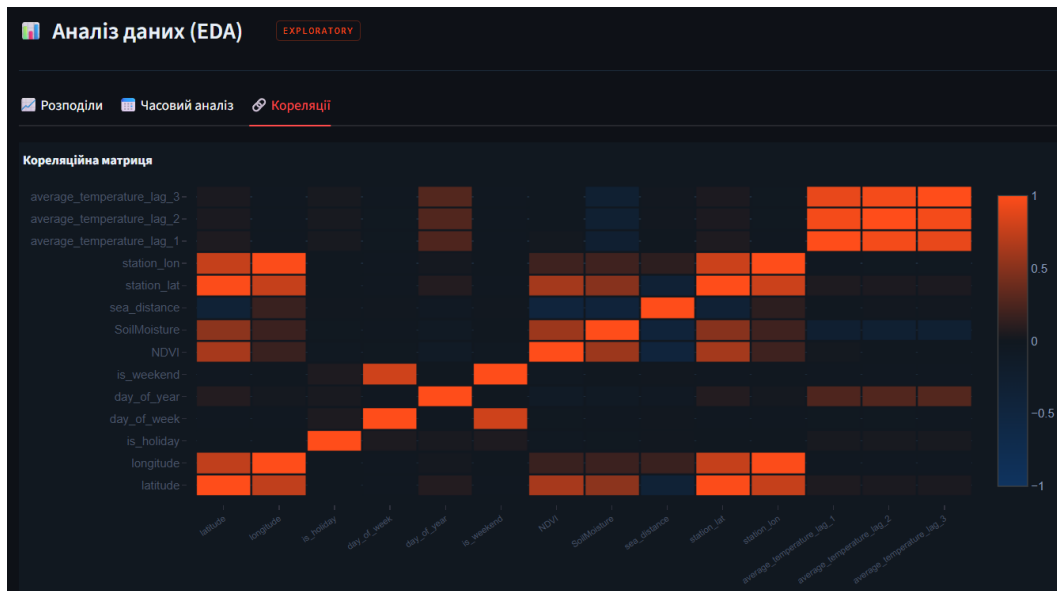


Рисунок 4.3 – Кореляційна матриця ключових числових ознак датасету

Візуалізація розподілів дозволяє виявити ключові відмінності між класами: зокрема, спостереження з пожежами ($is_fire = 1$) характеризуються систематично нижчими значеннями NDVI та вологості ґрунту, що узгоджується з фізичними закономірностями, описаними в підрозділі 1.1.

Аналіз гістограми розподілу NDVI підтверджує статистично значущу різницю між класами, що свідчить про високу дискримінаційну здатність цієї ознаки та обґрунтовує її включення до складу вхідного вектора моделі.

Виражена сезонність пожежних подій, зафіксована у розподілі по місяцях, підтверджує необхідність включення часових ознак (місяць року) до навчальної вибірки та обґрунтовує вибір стратифікованого розбиття даних зі збереженням часової репрезентативності.

4.3 Реалізація пайплайну передобробки даних

Завантаження датасету реалізовано через клас-завантажувач, що виконує послідовну валідацію наявності обов'язкових стовпців, перетворення типу даних дати, видалення 30 нерелевантних лагових ознак координат (артефакт вихідної структури даних) та формування стратифікованої вибірки із збереженням балансу класів через окрему вибірку для кожного класу з подальшим об'єднанням.

Пайплайн передобробки реалізовано як послідовність окремих трансформерів, що відповідають інтерфейсу `scikit-learn` (методи `fit/transform`), але керуються явно (без використання вбудованого класу `Pipeline`) через виявлену проблему серіалізації – десеріалізований через `joblib` об'єкт `Pipeline` генерував помилку `NotFittedError` при викликах `transform()` у `production`-середовищі. Реалізовано п'ять послідовних трансформерів: імпутація пропущених значень медіаною кожної ознаки; вінсоризація (обрізка значень за межами 1-го та 99-го перцентилів навчальної вибірки); логарифмічне перетворення скошених ознак (опад, сніг, туман, грім) функцією $\log(1+x)$; вибір фінального набору колонок моделі; опціональна стандартизація ознак (застосовується лише для `Logistic Regression` та `SVM`, оскільки дерево-базовані моделі нечутливі до масштабу ознак).

Реалізовано обчислення п'яти синтетичних ознак, формули яких наведено в підрозділі 3.4.2: температурного діапазону доби, мультиплікативної взаємодії температури та вітру, індексу посушливості, бінарного прапора аномалії `NDVI` (поріг – 15-й перцентиль навчальної

вибірки) та комплексного індексу погодного пожежного ризику (рисунок 4.4).

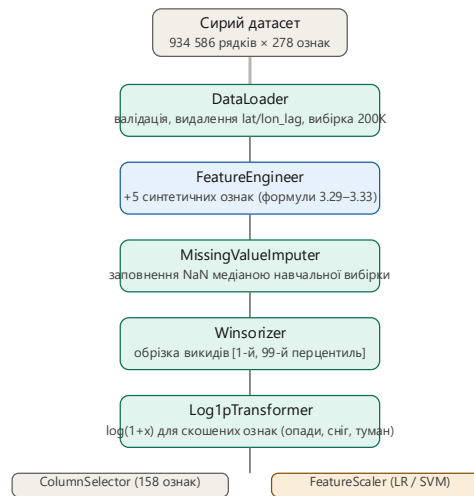


Рисунок 4.4 – Схема послідовності трансформерів пайплайну передобробки даних

Таким чином, розроблений пайплайн забезпечує повну відтворюваність трансформацій даних від вихідного формату до фінального вхідного вектора моделі, що є необхідною умовою коректності порівняльних експериментів.

4.4 Навчання та порівняльне оцінювання моделей

Дані розбито на навчальну (70%, 140 000 спостережень), валідаційну (10%, 20 000) та тестову (20%, 40 000) вибірки методом стратифікованого розбиття із збереженням балансу класів у кожній підмножині та фіксованим параметром відтворюваності. Модель SVM через квадратичну обчислювальну складність навчена на репрезентативній підвибірці 20 000 спостережень замість повної навчальної вибірки.

4.4.1 Результати навчання моделей для регіону Марокко

Модель XGBoost обрано основною («production») моделлю системи завдяки найвищим показникам якості за всіма метриками (таблиця 4.3) [1]. Матриця помилок XGBoost на тестовій вибірці: $TN = 18686$, $FP = 1314$, $FN = 617$, $TP = 19383$ – отримане співвідношення $FN < FP$ свідчить про «обережний» характер моделі (рисунки 4.5, 4.6), що є бажаною властивістю для задач превентивного моніторингу, де пропущена пожежа (FN) є небезпечнішою помилкою, ніж хибна тривога (FP).

Таблиця 4.3 – Порівняльна таблиця метрик якості чотирьох моделей (Марокко, тестова вибірка $n = 40\,000$)

Модель	Accuracy	Precision	Recall	F1	ROC-AUC	LogLoss
XGBoost	0.952	0.936	0.969	0.953	0.991	0.142
Random Forest	0.915	0.886	0.954	0.919	0.977	0.234
SVM (RBF)	0.854	0.825	0.900	0.861	0.921	0.351
Logistic Regression	0.820	0.792	0.869	0.829	0.897	0.397

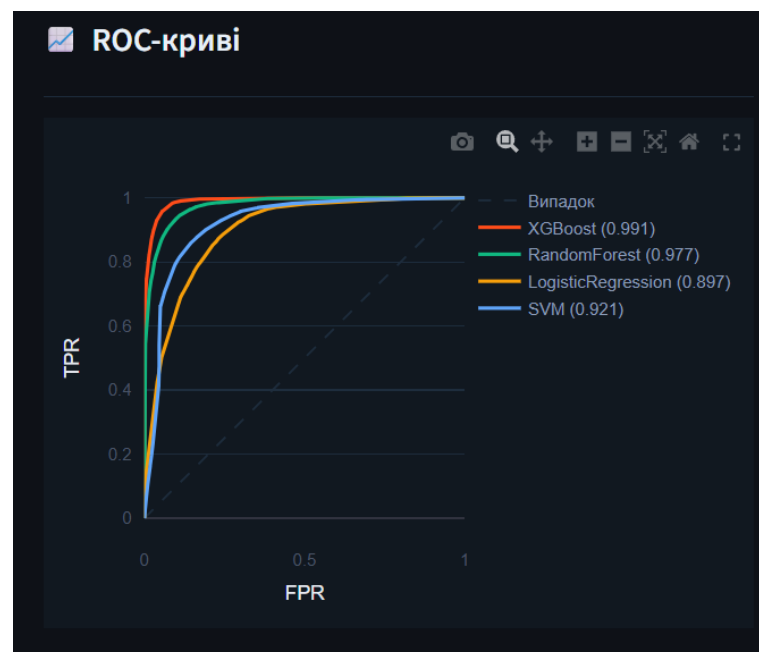


Рисунок 4.5 – ROC-криві чотирьох моделей (Марокко) на одному графіку

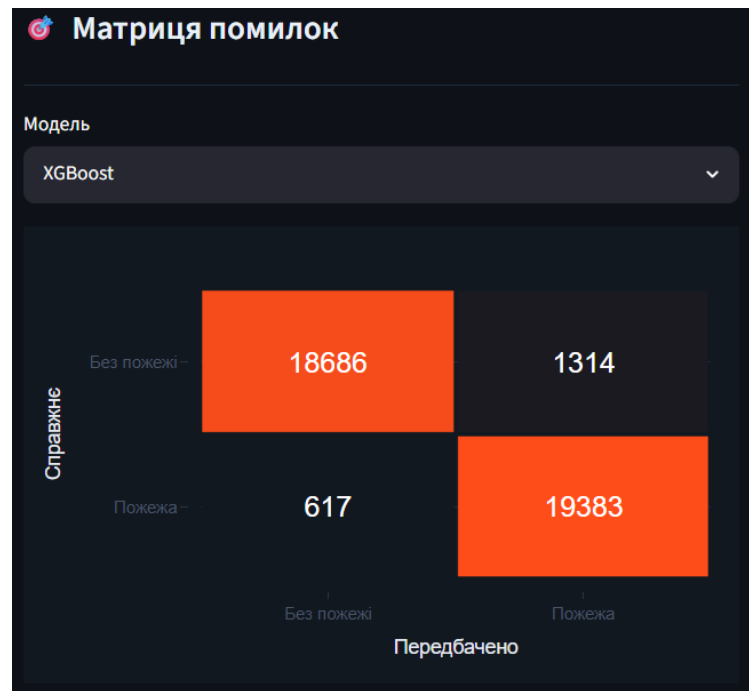


Рисунок 4.6 – Матриця помилок моделі XGBoost (теплова карта)

4.4.2 Результати навчання моделей для регіону Алжир

Результати навчання підтверджують теоретичне припущення про перевагу ансамблевих методів над лінійними моделями при роботі з нелінійно взаємодіючими екологічними предикторами, що узгоджується з аналізом, проведеним у розділі 2 (таблиця 4.4).

Таблиця 4.4 – Порівняльна таблиця метрик якості моделей (Алжир, тестова вибірка n=49)

Модель	Accuracy	Precision	Recall	F1	ROC-AUC	LogLoss
Random Forest	0.959	0.933	1.000	0.966	1.000	0.080
Logistic Regression	0.939	0.931	0.964	0.947	0.993	0.150

Отримане значення AUC=1.000 для Random Forest на алжирському датасеті потребує критичної інтерпретації (таблиця 4.4), детально обговореної в підрозділі 4.7 [2]: малий обсяг тестової вибірки (49

спостережень) у поєднанні із сильною природною кореляцією компонентів FWI із цільовою змінною (оскільки FWI історично розроблявся спеціально для оцінювання пожежної небезпеки) може призводити до оптимістично завищеної оцінки якості моделі, що не обов'язково узагальнюється на ширшу вибірку (рисунки 4.7, 4.8).



Рисунок 4.7 – ROC-криві моделей Random Forest та Logistic Regression (Алжир)

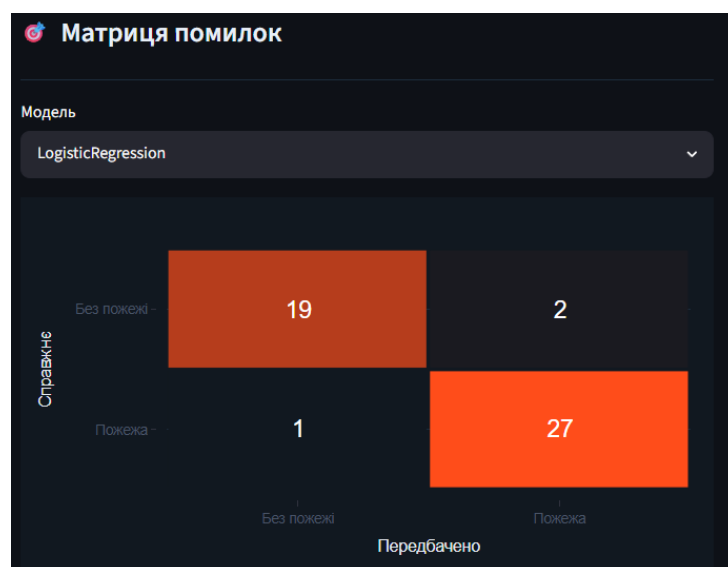


Рисунок 4.8 – Матриця помилок моделі Logistic Regression (Алжир)

4.5 Реалізація модуля пояснювального ШІ

Початкова реалізація передбачала обчислення SHAP-значень безпосередньо під час роботи вебзастосунку через виклик TreeExplainer у реальному часі. Практичне тестування виявило проблему – одночасне утримання в пам'яті повного датасету, навченої моделі та екземпляра пояснювача. Архітектурним рішенням стало перенесення обчислення глобальних SHAP-значень на етап навчання моделі (виконується один раз на репрезентативній вибірці зі 100 тестових спостережень) із подальшим кешуванням результату в JSON-файл, який вебзастосунок лише завантажує без повторного звернення до бібліотеки SHAP.

4.5.1 Глобальна важливість ознак та інтерпретація (Марокко)

Було здійснено груповий розподіл важливості різних ознак: Geographic – 45%, Temperature – 25%, Satellite – 13%, Wind – 9%, Precipitation – 7%, Engineered – 0% (рисунок 4.10). Сформовано таблицю найвпливовіших ознак моделі XGBoost відповідно до SHAP-аналізу (таблиця 4.5).

Таблиця 4.5 – Список 10 найважливіших ознак моделі XGBoost за результатами SHAP-аналізу

Ранг	Ознака	Важливість (mean)	Категорія
1	Latitude	1.192	Geographic
2	NDVI	0.791	Satellite
3	Sea Distance	0.661	Geographic
4	Longitude	0.335	Geographic
5	Station Lon	0.227	Geographic
6	Station Lat	0.224	Geographic
7	Dew Point (D-1)	0.163	Temperature
8	Max Temperature (D-1)	0.149	Temperature
9	SoilMoisture	0.112	Satellite
10	Precipitation Yearly Mean	0.080	Precipitation

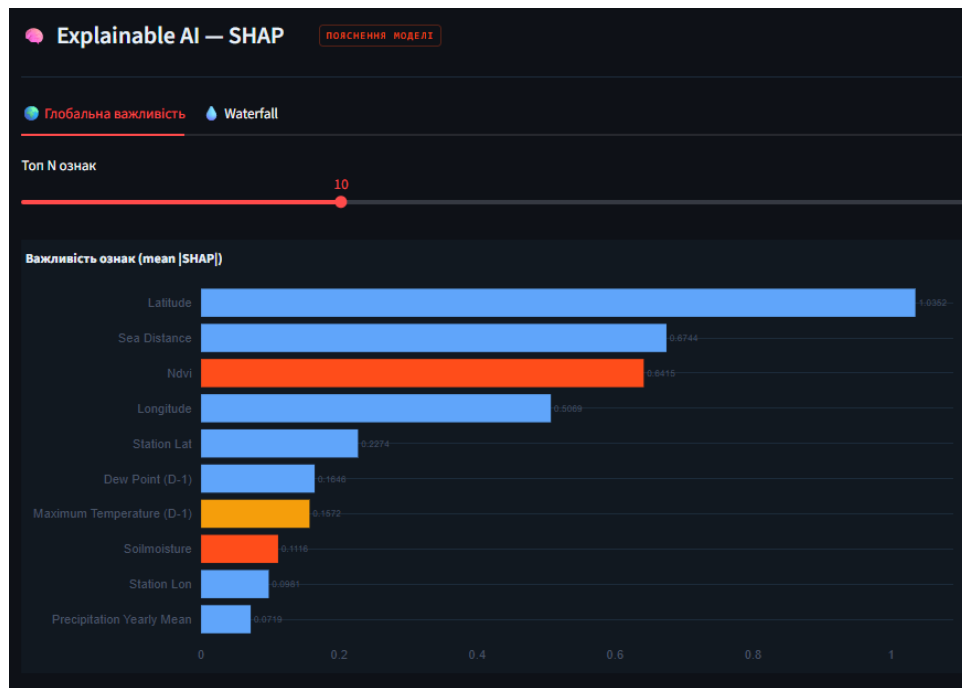


Рисунок 4.9 – Стовпчикова діаграма 10 ознак за важливістю SHAP



Рисунок 4.10 – Груповий розподіл важливості ознак за категоріями

Отримане домінування координат (44% сумарної важливості) пояснюється тим, що в територіально неоднорідному регіоні Марокко географічні координати виступають проксі-змінною кліматичної зони (теоретично обґрунтовано в підрозділі 2.3.3): північні вологіші регіони та південні аридні передсахарські території мають принципово різний базовий рівень пожежного ризику, і модель самостійно виявила цю

закономірність без явного програмування. Це водночас є методологічним застереженням: модель де-факто є регіонально специфічною і потребує повторного навчання для застосування до територій з іншою кліматичною зональністю.

Прийняте архітектурне рішення щодо попереднього обчислення та кешування SHAP-значень забезпечило зниження часу відгуку інтерфейсу модуля пояснювального аналізу з десятків секунд до мілісекунд, що є критичним для комфортної інтерактивної роботи користувача.

4.5.2 Глобальна важливість ознак (Алжир)

Для допоміжної моделі Random Forest на датасеті Алжиру найважливішими ознаками виявилися компоненти ISI (33%) та FFMC (30%), на відміну від моделі Марокко, де домінують географічні та супутникові ознаки (рисунок 4.11). Це пояснюється відсутністю в алжирському датасеті геопросторових і супутникових предикторів – модель спирається безпосередньо на вже агреговані метеорологічні індекси системи FWI.

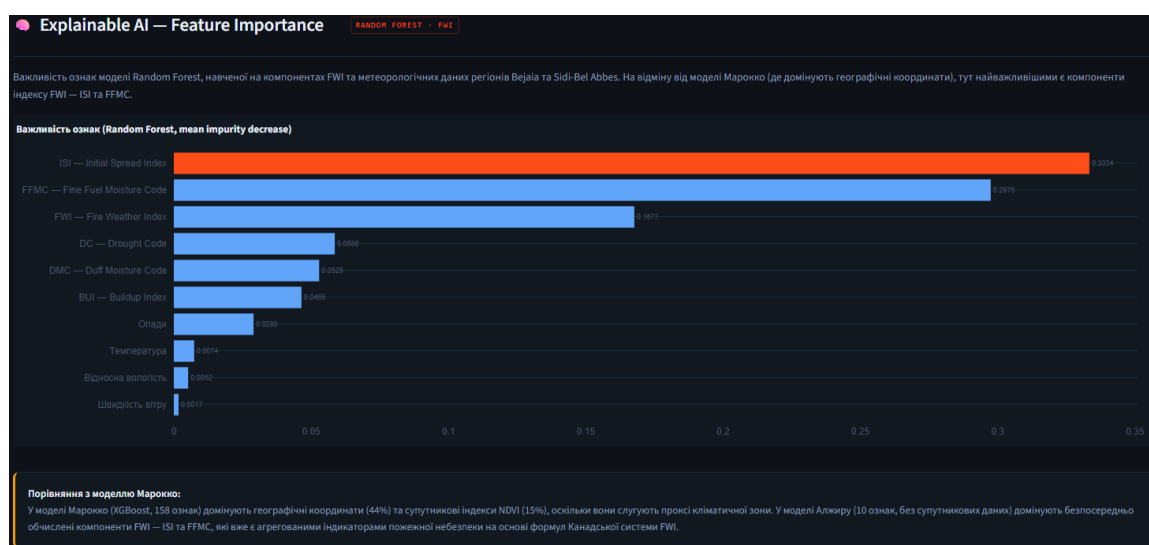


Рисунок 4.11 – Важливість ознак моделі Random Forest (Алжир)

4.5.3 Локальне пояснення (waterfall)

Для одного конкретного прогнозу реалізовано waterfall-візуалізацію, що декомпозиє прогнозовану ймовірність на суму внесків окремих ознак відносно базового значення моделі, з дотриманням властивості локальної точності SHAP (рисунок 4.12).

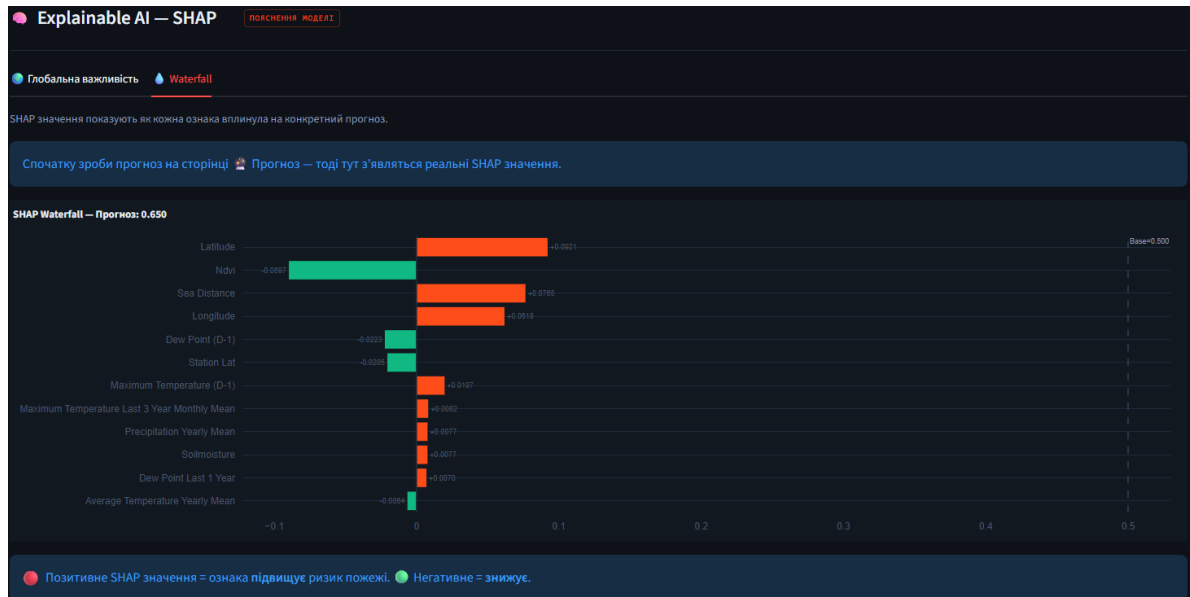


Рисунок 4.12 – Waterfall-діаграма для конкретного прогнозу ризику

4.6 Опис реалізованого вебзастосунок

4.6.1 Загальна структура інтерфейсу

Вебзастосунок реалізовано на базі фреймворку Streamlit як багатосторінковий дашборд із дев'ятьма функціональними сторінками та перемикачем регіону (Марокко / Алжир) у бічній навігаційній панелі (таблиця 4.6).

Таблиця 4.6 – Опис сторінок вебзастосунку FFIS

№	Сторінка	Призначення
1	Головна	KPI-показники, сезонний розподіл, рейтинг моделей
2	Прогноз	Інтерактивне прогнозування ризику в реальному часі для обраного регіону
3	Карта	Геопросторова теплова карта, маркери, супутникові шари NASA GIBS
4	Аналітика	Розвідувальний аналіз даних (EDA)
5	Порівняння	Таблиця метрик, ROC-криві, матриці помилок
6	SHAP / XAI	Глобальна та локальна інтерпретація моделі
7	Сценарії	Симуляція «що якщо», топ-зони ризику, авто-звіт
8	Датасет	Схема, статистика, вибірка з можливістю експорту
9	Про проєкт	Технічний стек, результати, перелік джерел

Головна сторінка вебзастосунку (рисунок 4.13) має такий вигляд.

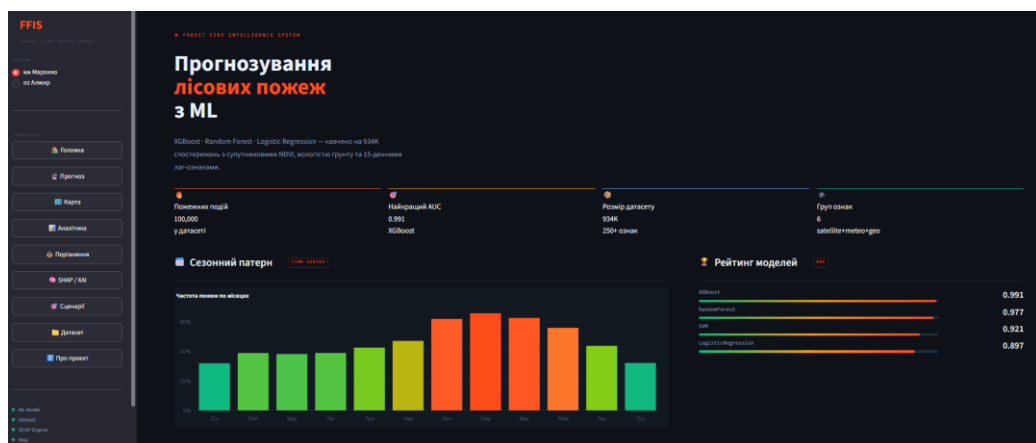


Рисунок 4.13 – Головна сторінка вебзастосунку FFIS

4.6.2 Модуль прогнозування в реальному часі

Для регіону Марокко користувач задає сім параметрів (координати, NDVI, вологість ґрунту, температура, опади, вітер), на основі яких система формує повний 158-вимірний вхідний вектор, заповнюючи відсутні лагові ознаки (дні 2–7) копією значення дня 1, а ковзні середні – тим самим значенням (припущення стабільності умов протягом тижня для одноразового прогнозу). Для регіону Алжир користувач безпосередньо

задає чотири метеопоказники та шість компонентів FWI, що передаються в модель Random Forest без додаткових перетворень (рисунки 4.14, 4.15).

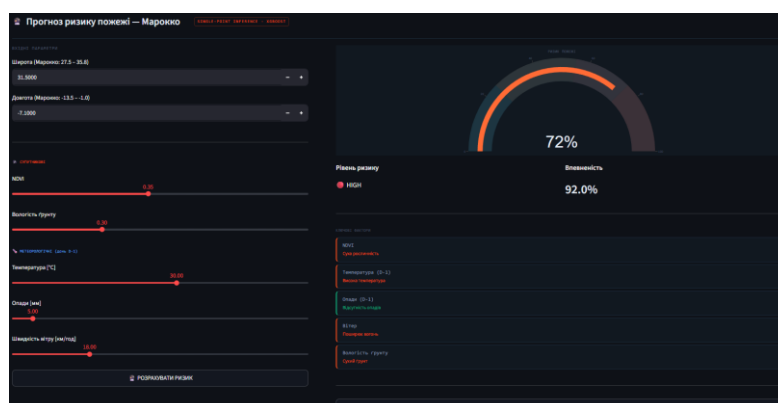


Рисунок 4.14 – Сторінка прогнозування для регіону Марокко (gauge-діаграма ризику, ключові фактори)

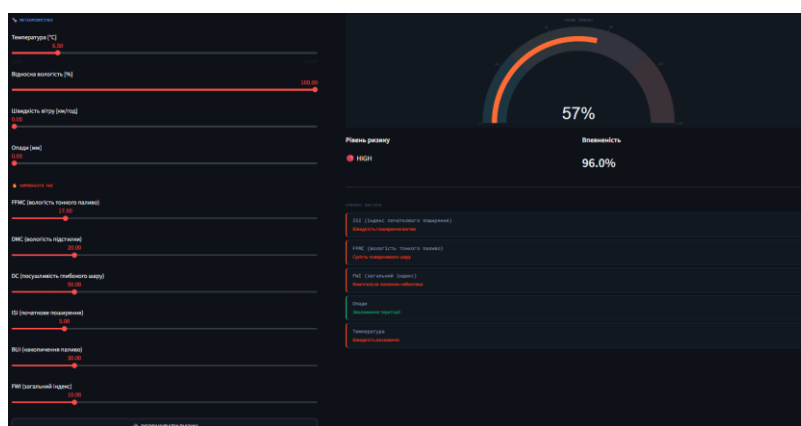


Рисунок 4.15 – Сторінка прогнозування для регіону Алжир

4.6.3 Геопросторова карта

Карта реалізована на основі бібліотеки Folium з кількома незалежно керуваними шарами: тепла карта прогнозованого ризику (плагін HeatMap), кластеризовані маркери пожеж (плагін MarkerCluster) з кольоровим кодуванням за рівнем ризику, шар «безпечних зон», а також інтегровані супутникові шари NASA GIBS (MODIS True Color, NDVI 8-Day

Composite) через протокол WMTS [17]. Для регіону Алжир, де відсутні точкові геокоординати спостережень, реалізовано спрощену карту з двома маркерами центрів провінцій, забарвленими відповідно до частки пожежних днів у кожному регіоні (рисунки 4.16, 4.17).



Рисунок 4.16 – Інтерактивна карта ризиків для регіону Марокко з тепловою картою та шаром NASA GIBS

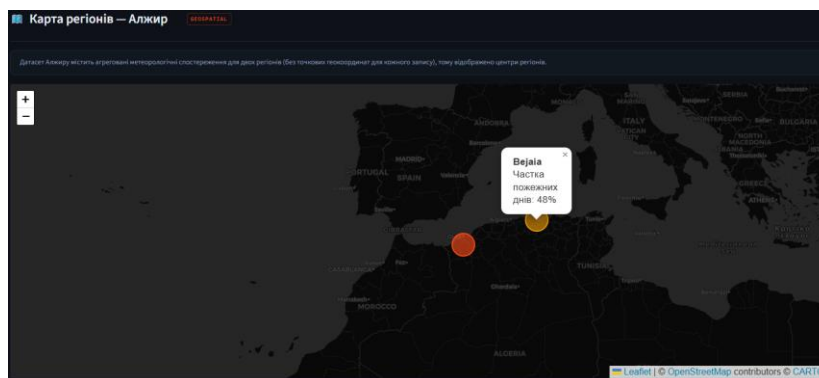


Рисунок 4.17 – Карта регіонів для Алжиру

4.6.4 Модуль AI-рекомендацій

Реалізовано додатковий модуль генерації текстових рекомендацій для служб пожежної безпеки на основі результату прогнозу. Основний шлях реалізації використовує локальну мовну модель (LLM) llama3.2:3b, що виконується через Ollama без потреби в зовнішньому платному API та без

передачі даних за межі локального середовища. Промпт для моделі сформульований англійською мовою з явними інструкціями щодо пропорційності рекомендацій рівню ризику (LOW / MEDIUM / HIGH / CRITICAL), що підвищило якість і релевантність відповідей порівняно з початковими спробами україномовного промπτу (рисунок 4.18). У разі недоступності локальної LLM передбачено детермінований резервний механізм – експертну систему на основі правил, що формує структуровані рекомендації на основі порогових значень ключових параметрів (NDVI, вологість ґрунту, температура, опади, вітер), гарантуючи безвідмовну роботу функції незалежно від доступності зовнішніх обчислювальних ресурсів.

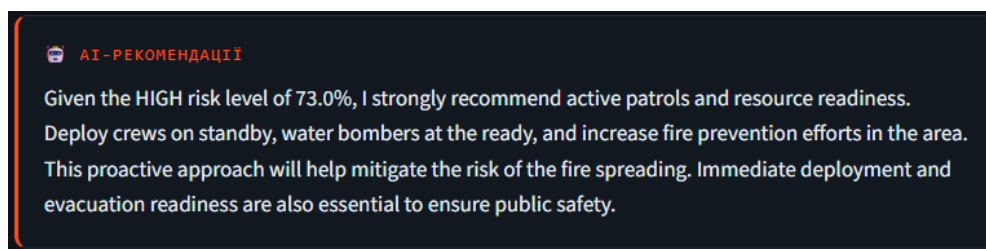


Рисунок 4.18 – Приклад згенерованих AI-рекомендацій для сценарію високого ризику

4.6.5 Модуль сценарної симуляції та звітності

Реалізовано додаткову сторінку «Сценарії» з трьома функціональними блоками: симуляцію впливу зміни окремих метеорологічних параметрів (підвищення температури, випадання опадів, посилення вітру, погіршення стану рослинності) на прогнозований ризик відносно базових умов із візуалізацією у вигляді порівняльної діаграми (рисунок 4.19); ранжування топ-N точок датасету з найвищим прогнозованим ризиком для пріоритетного патрулювання; автоматичну

генерацію текстового зведеного звіту про поточну ситуацію в регіоні з можливістю завантаження (рисунок 4.20).



Рисунок 4.19 – Сторінка симуляції сценаріїв з порівняльною діаграмою

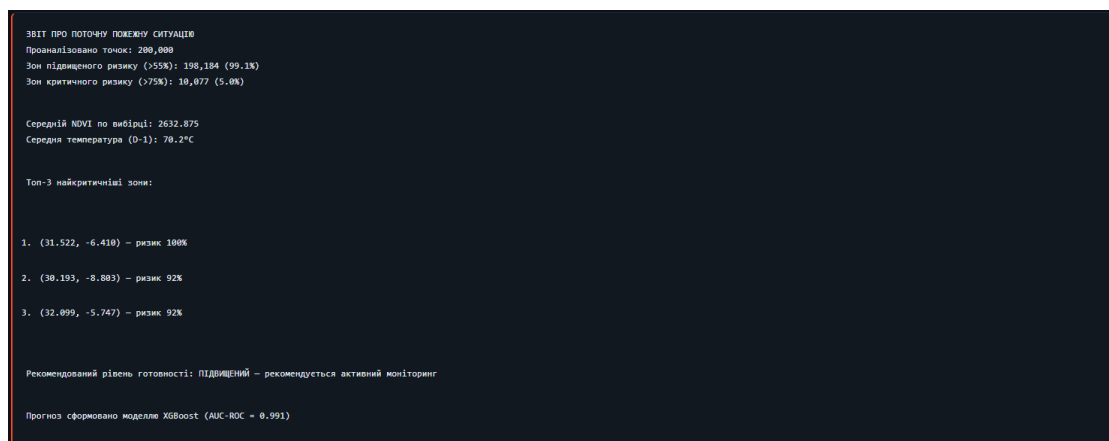


Рисунок 4.20 – Автоматично згенерований текстовий звіт

Модуль сценарної симуляції реалізує концепцію «what-if» аналізу, дозволяючи користувачу досліджувати чутливість прогнозованого ризику до зміни окремих метеорологічних параметрів та візуально порівнювати результати кількох сценаріїв на одному графіку.

Порівняльна діаграма сценаріїв наочно демонструє нелінійний характер залежності прогнозованого ризику від зміни вхідних параметрів, що підтверджує обґрунтованість використання ансамблевих моделей замість лінійних для задач подібного класу.

4.7 Тестування системи та обговорення результатів

4.7.1 Тестування коректності прогнозування та обмеження системи

Проведено систематичне тестування чутливості моделі XGBoost до зміни кожного з семи вхідних параметрів при фіксованих інших значеннях. Підтверджено фізично коректну монотонну реакцію моделі для шести з семи параметрів (зростання ризику при зниженні NDVI/вологості ґрунту/опадів та підвищенні температури/вітру). Виявлено локальну розривність прогнозу в околі нульової швидкості вітру – артефакт розподілу навчальних даних, де майже відсутні спостереження з точно нульовим вітром.

Виявлено й чесно задокументовано чотири основні обмеження розробленої системи. По-перше, п'ять синтетичних інженерних ознак моделі Марокко отримали нульову SHAP-важливість для XGBoost, оскільки ансамбль дерев здатний самостійно відтворити еквівалентні нелінійні комбінації з базових ознак – це не свідчить про методологічну помилку, а ілюструє різницю в потребі явної інженерії ознак між лінійними та ансамблевими моделями. По-друге, висока важливість географічних координат означає регіональну специфічність моделі Марокко. По-третє, AUC = 1.000 моделі Random Forest на датасеті Алжиру через малий обсяг тестової вибірки (49 спостережень) потребує обережної інтерпретації та подальшої валідації на ширших даних. По-четверте, формування лагових ознак для одноразового прогнозу користувача спирається на спрощене припущення стабільності погодних умов протягом тижня.

4.7.2 Порівняльний аналіз двох регіональних моделей

Порівняння (таблиця 4.7) демонструє принципову відмінність підходів: за наявності багатого, об'ємного датасету із супутниковими індексами модель самостійно виявляє складні географічно зумовлені закономірності (Марокко); за обмеженого обсягу даних без супутникових індексів ефективнішим виявляється спирання на вже агреговані, фізично обґрунтовані індекси на кшталт системи FWI (Алжир), хоча отримана оцінка якості у такому випадку є менш статистично надійною через малий розмір вибірки.

Таблиця 4.7 – Порівняльна характеристика моделей для двох досліджуваних регіонів

Критерій	Марокко (XGBoost)	Алжир (Random Forest)
Обсяг навчальних даних	140 000	194
Кількість ознак	158	10
Тип предикторів	Геопросторові + супутникові + метео	Метеорологічні + FWI
ROC-AUC	0.9905	1.000
Домінуючий фактор	Географічні координати (44%)	ISI, FFMC (63% разом)
Надійність оцінки	Висока (велика тестова вибірка)	Потребує додаткової валідації

Цей висновок має визначальне значення для стратегії подальшого масштабування системи. В умовах реальної експлуатації служби моніторингу регулярно стикаються з проблемою неоднорідності даних. Для одних регіонів може бути доступний повний спектр супутникових знімків, тоді як для інших (унаслідок щільної хмарності чи задимлення) доводиться спиратися виключно на локальні вимірювання наземних метеостанцій. З огляду на це, архітектура системи доводить свою гнучкість та адаптивність.

ВИСНОВКИ

У кваліфікаційній роботі вирішено актуальну науково-прикладну задачу – розроблено інтелектуальну інформаційну систему прогнозування ризику виникнення лісових пожеж на основі методів машинного навчання, що охоплює два географічні регіони Північної Африки та реалізована у вигляді завершеного інтерактивного вебзастосунку. На основі виконаного дослідження сформульовано такі висновки.

У межах аналізу предметної галузі встановлено, що лісові пожежі є складними фізико-екологічними явищами, динаміка яких визначається взаємодією метеорологічних, топографічних факторів та характеристик горючого матеріалу. Регіон Північної Африки (гірські масиви Атласу в Марокко та Алжирі) обрано як репрезентативний об'єкт дослідження через унікальне поєднання аридного клімату, специфічних вітрових режимів та займистої ендемічної рослинності. Проведено критичний аналіз існуючих систем моніторингу (EFFIS, GWIS, NASA FIRMS) та виявлено спільне обмеження – відсутність проактивного прогнозування з адаптивними алгоритмами машинного навчання та механізмами пояснення рішень.

На теоретичному рівні систематизовано математичний апарат чотирьох алгоритмів бінарної класифікації (логістична регресія, метод опорних векторів з RBF-ядром, Random Forest на основі бутстреп-агрегації, XGBoost з оптимізацією через розклад Тейлора другого порядку) та методу пояснювального штучного інтелекту SHAP, що базується на теорії значень Шеплі з кооперативних ігор. Теоретично обґрунтовано доцільність застосування лагових метеорологічних ознак (1–7 днів) та агрегованих часових статистик як інформативних предикторів кумулятивного дефіциту вологи в рослинності та ґрунті.

У межах практичної реалізації навчено чотири класифікаційні моделі на стратифікованій вибірці обсягом 200 000 спостережень (Марокко) та дві моделі на повному датасеті (Алжир). Найвищу якість продемонстрував

алгоритм XGBoost: $AUC-ROC = 0.9905$, $F1 = 0.953$, $Accuracy = 0.952$, $Recall = 0.969$ на тестовій вибірці з 40 000 спостережень, що підтверджує теоретичне припущення про перевагу градієнтного бустингу над лінійними та ядровими методами для даних з нелінійною структурою та мультиколінеарними предикторами.

За результатами SHAP-аналізу встановлено, що домінуючими предикторами моделі XGBoost є географічні координати (44% сумарної важливості), вегетаційний індекс NDVI та показник вологості ґрунту. Виявлено та теоретично інтерпретовано феномен географічних координат як проксі-змінної кліматичної зони в умовах просторово неоднорідного регіону. Для моделі Алжиру домінування компонентів FWI (ISI – 33%, FFMC – 30%) підтверджує фізичну обґрунтованість Канадської системи індексів пожежної небезпеки для прогнозування в середземноморських регіонах.

Розроблено дев'ятисторінковий інтерактивний вебзастосунок на базі Streamlit, що включає: модуль прогнозування в реальному часі для двох регіонів через перемикач Марокко/Алжир; геопросторову карту з інтеграцією супутникових шарів NASA GIBS [17]; сторінку розвідувального аналізу даних; порівняння моделей з ROC-кривими та матрицями помилок; SHAP-аналіз з глобальним та локальним (waterfall) поясненням; сторінку сценарної симуляції та автоматичної звітності; модуль AI-рекомендацій на основі локальної мовної моделі з детермінованим резервним механізмом.

Перспективами подальшого розвитку системи є: інтеграція з реальними метеорологічними API для автоматичного формування лагових ознак; розширення бази даних для інших країн Середземного моря, Захід США та південні регіони Австралії для підвищення надійності статистичного оцінювання якості моделі; реалізація автоматичного перенавчання моделей при надходженні нових підтверджених спостережень; розширення системи на інші пожежонебезпечні регіони Середземноморського басейну – Іспанію, Грецію, Туреччину.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

- 1) Chen T., Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA, 2016. P. 785–794. URL: <https://doi.org/10.1145/2939672.2939785> (date of access: 14.06.2026).
- 2) Breiman L. Random forests. *Machine Learning*. 2001. Vol. 45, no. 1. P. 5–32. URL: <https://doi.org/10.1023/A:1010933404324> (date of access: 14.06.2026).
- 3) Cortes C., Vapnik V. Support-vector networks. *Machine Learning*. 1995. Vol. 20, no. 3. P. 273–297. URL: <https://doi.org/10.1007/BF00994018> (date of access: 14.06.2026).
- 4) Cox D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B*. 1958. Vol. 20, no. 2. P. 215–242. URL: <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x> (date of access: 14.06.2026).
- 5) Shapley L. S. A value for n-person games. *Contributions to the Theory of Games*. Princeton, 1953. Vol. 2. P. 307–317.
- 6) Monitoring vegetation systems in the Great Plains with ERTS / J. W. Rouse et al. *Third ERTS Symposium*. Washington, 1974. P. 309–317.
- 7) The MODIS fire products / C. O. Justice et al. *Remote Sensing of Environment*. 2002. Vol. 83, no. 1-2. P. 244–262. URL: [https://doi.org/10.1016/s0034-4257\(02\)00076-7](https://doi.org/10.1016/s0034-4257(02)00076-7) (date of access: 14.06.2026).
- 8) Giglio L., Schroeder W., Justice C. O. The collection 6 MODIS active fire detection algorithm and fire products. *Remote Sensing of Environment*. 2016. Vol. 178. P. 31–41. URL: <https://doi.org/10.1016/j.rse.2016.02.054> (date of access: 14.06.2026).

9) Abid F. A survey of machine learning algorithms based forest fires prediction and detection systems. *Fire Technology*. 2021. Vol. 57. P. 559–590. URL: <https://doi.org/10.1007/s10694-020-01056-z> (date of access: 14.06.2026).

10) Development of a framework for fire risk assessment using remote sensing and geographic information system technologies / E. Chuvieco et al. *Ecological Modelling*. 2010. Vol. 221, no. 1. P. 46–58. URL: <https://doi.org/10.1016/j.ecolmodel.2008.11.017> (date of access: 14.06.2026).

11) A review of machine learning applications in wildfire science and management / P. Jain et al. *Environmental Reviews*. 2020. Vol. 28, no. 4. P. 478–505. URL: <https://doi.org/10.1139/er-2020-0019> (date of access: 14.06.2026).

12) Sayad Y. O., Mousannif H., Al Moatassime H. Predictive modeling of wildfires: a new dataset and machine learning approach. *Fire Safety Journal*. 2019. Vol. 104. P. 130–146. URL: <https://doi.org/10.1016/j.firesaf.2019.01.006> (date of access: 14.06.2026).

13) European forest fire information system (EFFIS). *Joint Research Centre, European Commission*. URL: <https://effis.jrc.ec.europa.eu> (date of access: 14.06.2026).

14) NASA-firms. *NASA | LANCE | FIRMS*. URL: <https://firms.modaps.eosdis.nasa.gov> (date of access: 14.06.2026).

15) GWIS - Welcome to GWIS. *GWIS - Welcome to GWIS*. URL: <https://gwis.jrc.ec.europa.eu> (date of access: 14.06.2026).

16) Algerian forest fires. *UCI Machine Learning Repository*. URL: <https://archive.ics.uci.edu/dataset/547/algerian+forest+fires+dataset> (date of access: 14.06.2026).

17) Global imagery browse services APIs | NASA earthdata. *NASA Earthdata*. URL: <https://earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/gibs> (date of access: 14.06.2026).

18) Climate change information for regional impact and for risk assessment. *Climate change 2021 – the physical science basis*. 2023. P. 1767–

1926. URL: <https://doi.org/10.1017/9781009157896.014> (date of access: 05.07.2026).

19) Global forest resources assessment 2020. *FAO*. URL: <https://doi.org/10.4060/ca8753en> (date of access: 14.06.2026).

20) Finney M. A. An overview of FlamMap fire modeling capabilities. *Fuels management – how much is enough? Proceedings of the RMRS-P-41*. Fort Collins, 2006. P. 213–220.