Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет *Комп'ютерних наук*
(повна назва)

Кафедра *Системотехніки*
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

(позначення документа)

Дослідження методів машинного навчання для прогнозування врожайності

(тема)

Виконала: студентка II курсу, групи *ІТПм-22-1*

Спеціальності *122 Комп'ютерні науки*
(код і повна назва спеціальності)

Тип програми освітньо-професійна

Освітня програма *Інформаційні технології проектування*
(повна назва освітньої програми)

Овчаренко А.Р.
(прізвище, ініціали)

Керівник *доц. Петрова Р.В.*
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри системотехніки _____ *проф. Гребеннік І.В.*
(підпис) (прізвище, ініціали)

2024 р.

Я, як студент ХНУРЕ, розумію і підтримую політику закладу із академічної доброчесності. Я не надавала і не одержувала недозволену допомогу під час підготовки кваліфікаційної роботи. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

«13» січня 2024 р.  *Овчаренко А.Р.*

Кваліфікаційна робота не містить відомостей заборонених до відкритого опублікування.

Кваліфікаційна робота виконана у відповідності до стандартів, що діють в Україні.

Попередній захист проведено «13» січня 2024 р.

Керівник кваліфікаційної роботи доц. Петрова Р.В 

Харківський національний університет радіоелектроніки

Факультет *Комп'ютерних наук*

Кафедра *Системотехніки*

Рівень вищої освіти *другий(магістерський)*

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

Тип програми *освітньо-професійна*

Освітня програма *Інформаційні технології проектування*

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___»_____20___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові *Овчаренко Аліні Ростиславівні*
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів машинного навчання для прогнозування врожайності
затверджена наказом по університету від 20.11 2023 р. № 1373Ст
2.Термін подання студентом роботи до екзаменаційної комісії 17.01.2024 р
3.Вихідні дані до роботи: *Дослідити методи машинного навчання у інформаційних системах прогнозування врожайності. Функції методу: обробка вхідних даних, складання прогнозу на основі методу випадкових лісів, візуалізація прогнозованих вихідних даних. Вибірки даних для навчання та тестування методу прогнозування. Операційна система Windows XP або вище, програмне забезпечення: CASE-засіб All Fusion Data Modeler (ERWin), All Fusion Process Modeler (BPWin), PyCharm (мова програмування Pyton).*
4.Перелік питань, що потрібно опрацювати в роботі *4.1 Вступ. 4.2 Аналіз предметної області. 4.3 Постановка задачі 4.4 Дослідження методів вирішення задачі 4.5 Експериментальні дослідження та перевірка результатів 4.6 Висновки.*
5.Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) *5.1 Концептуальна діаграма, 5.2 Декомпозиція концептуальної діаграми, 5.3 Невідформатований датасет*

5.4 Робота програмного засточунку 5.5 Підбір палітри кольорів, 5. 6 Інтерфейс програмного засобу, 5.7 Карта сайту.

6.Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	*Доц. Петрова Р.В.*		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів	Примітка
1	*Отримання завдання на виконання роботи*	*20.11.23*	*виконано*
2	*Огляд матеріалів та аналіз предметної області*	*23.11.23-30.11.23*	*виконано*
3	*Визначення досліджуваних методів машинного навчання*	*02.12.23*	*виконано*
4	*Дослідження методів машинного навчання в прогнозуванні*	*03.12.23-23.12.23*	*виконано*
5	*Розробка частини інформаційної системи*	*23.12.23-25.12.23*	*виконано*
6	*Визначення функцій інтерфейсу клієнтської частин інформаційної системи*	*27.12.23-30.12.23*	*виконано*
7	*Розробка елементів інтерфейсу*	*30.12.23*	*виконано*
8	*Оформлення пояснювальної записки*	*02.01.24-10.01.24*	*виконано*
9	*Оформлення додатків*	*11.01.24*	*виконано*
10	*Представлення на рецензування*	*14.01.24*	*виконано*

Дата видачі завдання 20.11. 2023 р.

Студент _____ *Овчаренко А.Р.*
(підпис)

Керівник роботи _____ *доц. Петрова Р.В.*
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи магістра містить: 92 с., 10 табл., 22 рис., 3 додатки, 32 джерел інформації.

ВРОЖАЙНІСТЬ, ВИПАДКОВІ ЛІСА, ДЕРЕВА РІШЕНЬ, ДОСЛІДЖЕННЯ, ГЛИБОКЕ НАВЧАННЯ, ІНФОРМАЦІЙНІ СИСТЕМИ, МАШИННЕ НАВЧАННЯ, ПРОГНОЗУВАННЯ.

Об’єктом дослідження є процес визначення прогнозів з використання машинного навчання на основі даних врожайності культур за останні роки.

Предметом дослідження є інформаційні технології й методи прогнозування даних.

Метою досліджень є застосування методів прогнозування даних через побудову численних дерев прийняття рішень для вирішення прикладних питань.

Методи дослідження – системний підхід, методи прогнозування, методи випадкових лісів, методи структурного аналізу і моделювання реляційних баз даних, підхід проектування програмного забезпечення.

Результат роботи – програмна реалізація методу прогнозування даних з врожайності культур, порівняння, дослідження та оцінка точності методу прогнозування.

Результати роботи можуть бути використані для проведення прогнозування даних, на основі яких користувач може сформувати рекомендації для себе що до посівів та догляду, а також може заздалегідь провести відповідну політику орієнтуючись на рівень ймовірного врожаю, уникаючи критичних збитків.

Область застосування – представлена система може бути використана на підприємствах аграрного сектору різних масштабів, окрім цього в ній будуть зацікавленні страхові кампанії та інвестори, що з ними контактують.

ABSTRACT

The explanatory note for the master's qualification work contains 92 pages, 10 tables, 22 figures, 3 appendices, 32 information sources.

YIELD, RANDOM FORESTS, DECISION TREES, RESEARCH, DEEP LEARNING, INFORMATION SYSTEMS, MACHINE LEARNING, FORECASTING .

The object of the research is the process of making predictions using machine learning based on crop yield data from recent years.

The subject of the research is information technologies and methods of data forecasting.

The research aims to apply data forecasting methods through the construction of numerous decision trees to address practical questions.

Research methods include a systemic approach, forecasting methods, random forest methods, structural analysis methods, modeling of relational databases, and software design approach.

The outcome of the work is the software implementation of a data forecasting method for crop yields, a comparison, research, and assessment of the accuracy of the forecasting method.

The results of the work can be used for data forecasting, based on which the user can generate recommendations for crop planting and care. Additionally, it enables the user to proactively implement a corresponding policy, guided by the anticipated level of harvest, thus avoiding critical losses.

Application area: the presented system can be employed in agricultural enterprises of various scales. Furthermore, insurance companies and interacting investors may find it of interest.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

CASE – сукупність методів і засобів автоматизованого проектування інформаційних систем; CASE-засоби BPWin ERWin, StarUML дозволяють проводити аналіз, документування і поліпшення бізнес процесів;

DFD – описуються потоки даних і дозволяється відстежувати обмін інформацією в системі між бізнес-процесами а також між системою і зовнішнім середовищем, їх можна використовувати щоб описати робочий процес і обробку інформації;

ID – ідентифікаційний номер в базі даних;

IDEF0 – модель функцій, методологія функціонального моделювання і графічного опису процесів, призначена для формалізації і опису бізнес-процесів. В IDEF0 розглядаються логічні зв'язки між роботами, а не послідовність їх виконання в часі;

SQL – Structured Query Language (мова структурованих запитів);

URL – Uniform Resource Locator (уніфікований покажчик інформаційного ресурсу);

Глибоке навчання - це підгалузь машинного навчання, яка використовує глибокі нейронні мережі для аналізу та розуміння складних завдань, таких як розпізнавання зображень та обробка мовлення.

ІС - інформаційна система;

Машинне навчання - це галузь штучного інтелекту, яка вивчає, як створювати алгоритми та моделі, які навчаються вирішувати завдання на основі даних, замість того, щоб бути явно програмованими.

ВСТУП

У сучасному світі збільшення продуктивності та ефективності сільського господарства є актуальною проблемою, особливо у контексті зміни клімату та загального росту населення. Наша країна має довгу традицію сільськогосподарського виробництва та є однією зі значущих аграрних держав у світі. Забезпечення стабільної врожайності є ключовим завданням для забезпечення продовольчої безпеки та економічного розвитку України.

Наявність точних та актуальних прогнозів врожаю є критично важливою для планування та оптимізації процесів сільського господарства. Дослідження та впровадження інноваційних методів для прогнозування врожаю та оптимізації виробництва стають все важливішими завданнями в цій галузі.

Магістерська робота присвячена розробці та впровадженню методів машинного навчання для прогнозування врожайності сільськогосподарських культур. Основною метою є створення інформаційної системи, яка дозволить аграрним підприємствам та фермерам отримувати точні та актуальні прогнози врожаю. Використання інноваційних підходів, таких як випадкові ліси та глибоке навчання, дозволить підвищити якість прогнозів та зменшити ризики в сільському господарстві.

Дана робота розглядає системний підхід до аналізу та моделювання даних врожайності, а також використання методів машинного навчання для побудови точних прогнозів. Проведені дослідження та практична реалізація системи прогнозування мають значущий потенціал для підвищення продуктивності та стійкості сільськогосподарського виробництва.

При підготовці кваліфікаційної роботи була підготовлена та опублікована наукова робота: «Дослідження Методів Машинного Навчання для Прогнозування» у рамках 12-ї Міжнародної науково-технічної конференції «Інформаційні системи та технології ІСТ-2023» у м. Харків, 28 листопада – 01 грудня 2023 р. Харків, Україна [1].

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Опис сучасного стану розвитку технологій машинного навчання

На сучасному етапі історії інформаційних технологій машинне навчання стало не тільки домінуючою областю досліджень, але й потужним інструментом, що надає можливість введення покращень у низці галузей, включаючи прогнозування врожайності. Завдяки активному розвитку технологій, доступності великих обсягів даних і зростаючій обчислювальній потужності, машинне навчання стає важливою ланкою для забезпечення точних та надійних прогнозів.

Машинне навчання [2] здобуло велику популярність завдяки своєму потенціалу вирішувати складні завдання без розгорнутого програмування. Алгоритми глибокого навчання та нейронні мережі здатні адаптуватися до різноманітних завдань, включаючи аналіз даних, класифікацію, регресію та генерацію контенту. Це стало можливим завдяки значному зростанню обчислювальної потужності та розвитку програмних інструментів.

Однією з найважливіших областей застосування машинного навчання є комп'ютерне бачення [3]. Нейронні мережі здатні до розпізнавання образів, що відкриває широкі можливості для автоматизації завдань, таких як розпізнавання обличь, автомобілів, об'єктів на медичних знімках та багато інших. Наприклад, системи комп'ютерного бачення застосовуються в одночасному розпізнаванні та підрахунку кількості плодів на деревах або оцінці стану рослин у сільському господарстві, що має велике значення для прогнозування врожайності.

Машинне навчання також знайшло застосування в обробці природних мов. Моделі глибокого навчання, такі як рекурентні нейронні мережі, здатні до автоматичного перекладу текстів з однієї мови на іншу, а також аналізу текстів для ефективного пошуку та витягнення інформації.

У галузі медицини машинне навчання стає важливим інструментом для діагностики та лікування різних захворювань, включаючи раннє прогнозування.

Автоматична обробка медичних зображень та аналіз даних допомагає лікарям швидко та точно визначити патологічні прояви.

У сільському господарстві, машинне навчання може застосовуватися для діагностики стану рослин та вчасного виявлення патологічних ознак, що впливає на якість та виживання культури. Не менш важливим є те що, ці технології можуть бути використані для аналізу великого обсягу інформації з різних джерел, що сприяє точному прогнозуванню можливої врожайності [4].

Рекомендаційні системи, побудовані на основі машинного навчання, стали важливими для сфери електронної комерції, а також для прогнозування споживчих вподобань та покупок. У сільському господарстві, ці системи можуть надавати рекомендації щодо оптимальних посівів, використання добрив та інших аспектів сільськогосподарської діяльності, що впливають на врожайність.

Машинне навчання грає важливу роль у сфері аналізу даних. Воно дозволяє виявляти залежності, визначати кореляції та робити прогнози на основі великих обсягів інформації. Такі можливості особливо важливі для вирішення завдань, пов'язаних із споживанням ресурсів, ефективністю та управлінням ризиками.

Розвиток машинного навчання суттєво змінює підходи до прийняття рішень в різних сферах, включаючи економіку, науку, медицину та сільське господарство. Сучасне суспільство вивчає можливості та виклики, пов'язані зі зростанням важливості машинного навчання у вирішенні проблем сьогодення.

Ці приклади підкреслюють важливість машинного навчання у прогнозуванні різних явищ, включаючи врожайність, та вказують на його значення для сільського господарства та сучасного суспільства в цілому.

1.2 Огляд проблеми сучасного стану прогнозування врожайності

Урожайність сільськогосподарських культур – ключовий показник, що визначає успішність аграрного виробництва. В Україні вирізняються дві ключові галузі – рослинництво та проміжна галузь, кормовиробництво. Ця проміжна

галузь має свою специфіку, структуру та організаційно-економічні основи, особливо у великих господарствах. Приблизно 93% орних земель в Україні відводиться під рослинництво та кормовиробництво, із них близько 30% призначено для вирощування кормових культур [4].

Ґрунтово-кліматичні умови України різноманітні в зонах Нечорноземної смуги, таких як Полісся, Лісостеп, та Степ. Кожна зона поділяється на північну, центральну і південну частини. Різні зони та підзони відрізняються ґрунтовим покривом, кількістю опадів, теплом, тривалістю вегетаційного періоду, і умовами перезимівлі. Врахування екологічних та біологічних особливостей цих умов є важливим для правильного розміщення сільськогосподарських культур в системі землекористування.

Зазначені раніше фактори, такі як різноманітні ґрунтово-кліматичні умови та гармонійне поєднання рослинництва та кормовиробництва, створюють в Україні сприятливі умови для успішного вирощування різноманітних сільськогосподарських культур. Враховуючи велику кількість орних земель, призначених для рослинництва та кормовиробництва, а також високий рівень біологічної орієнтованості в аграрному секторі, Україна стає ідеальним місцем для вирощування різноманітних культур.

Сприятливі умови для вирощування різних культур визначаються не лише розмаїттям ґрунтових та кліматичних умов у різних зонах країни, але і узгодженими методами вирощування, які сприяють максимальному використанню біологічних ресурсів та зниженню хімічного впливу. Це надає можливість аграрному сектору України не лише забезпечувати внутрішні потреби в продукції, а й ставати конкурентоспроможним гравцем на світовому ринку сільськогосподарської продукції.

В сучасному світі велика увага приділяється розробці та вдосконаленню методів прогнозування для оптимізації вирощування рослин і забезпечення стабільності агросектору. На даний момент існують різноманітні підходи до прогнозування врожайності [5], які охоплюють широкий спектр методів, від традиційних до сучасних.

Серед аналогів, CISS GROUP [6] пропонує простий інтерфейс, де можна залишити відповідний запит на обстеження для прогнозу врожайності. Вони переважно спрямовані на послуги з обслуговування сільськогосподарської техніки, а також надають оцінки земель і терміни дозрівання врожаю на основі системи моніторингу.

Однак існують суттєві недоліки, серед яких перш за все є те, що користувач послуги не може вибрати налаштування, або комплексно перевірити прогноз врожаю за тих чи інших умов. Окрім цього перевірка йде відносно існуючої ділянки, для котрої виділяється велика кількість замірів, що потребують постійного оновлення. Так як методика надання послуг чітко не зазначена, можна зробити припущення що надається оцінка врожаю методом експертних оцінок [5], що ґрунтується на досвіді та інтуїції спеціалістів. Однак цей метод має досить значну частку суб'єктивності, тому надійність його недостатня.

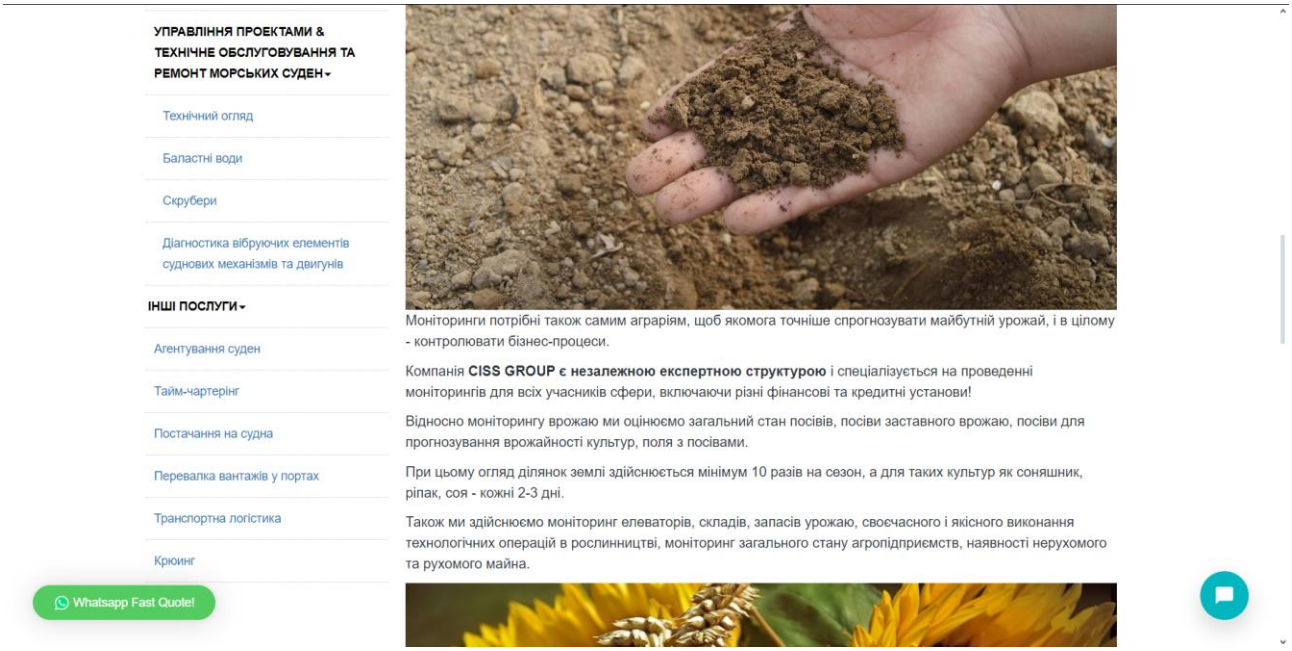


Рисунок 1.1 – Головна сторінка сайту «CISS GROUP»

EOS [7] має комплекс підходів для прогнозування, як на основі біофізичних показників, так і на основі статистики, однак, переважно покладається систему супутникового спостереження, що визначає стан полів

візуально. Вони володіють високою якістю прогнозів та використовують моделі машинного навчання для їх покращення. Однак, за для високої точності видачі прогнозу необхідно звертатись не раніше, ніж за 3 місяці до орієнтованого збору врожаю, що накладає суттєві обмеження.

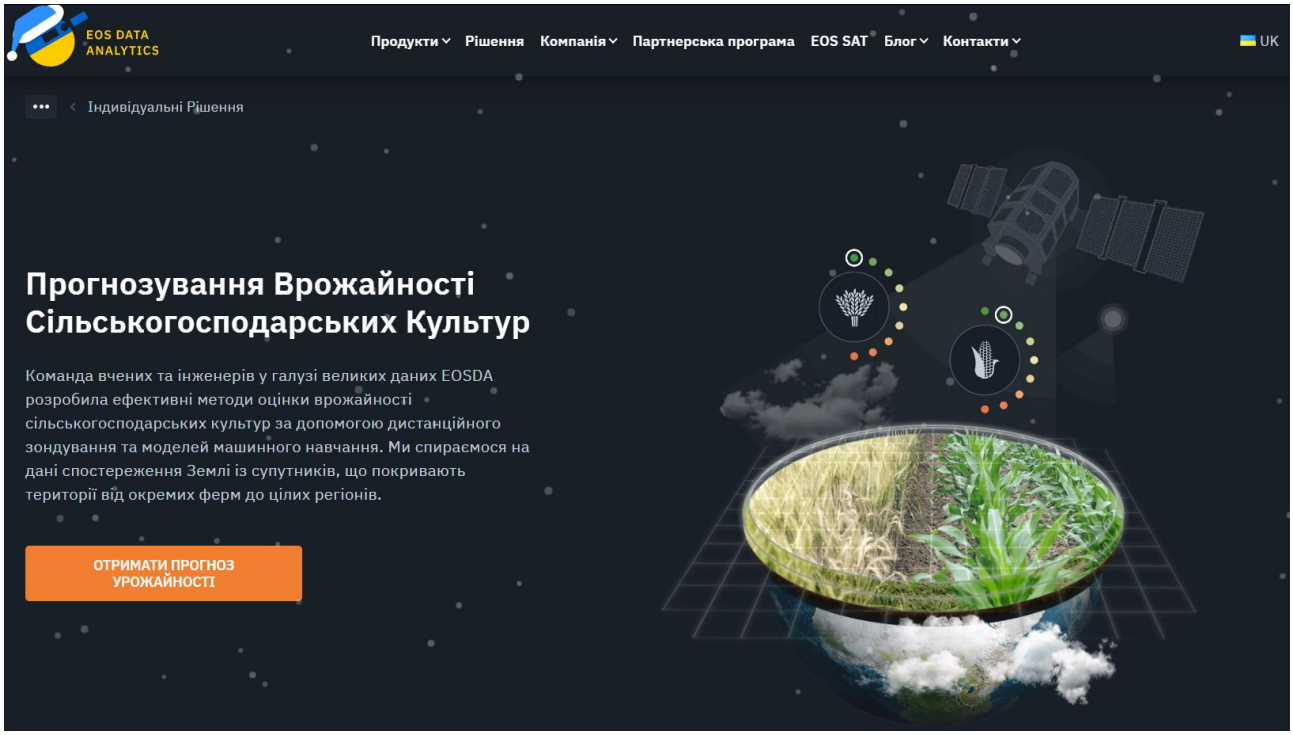


Рисунок 1.2 – Головна сторінка сайту «EOS»

Metos [8] надає послуги прогнозування врожайності на основі збору статистичних даних. Однак, вони більше орієнтовані на оцінку прибутковості товарів. Список прогнозованих культур обмежений, при необхідності доповнення ряду культур необхідно виконати ряд складних інструкцій. Для прогнозування використовуються статистичні данні, розрахунки йдуть відносно закону мінімуму [4], роблячи припущення, що найбільше бракуватиме саме опадів. Саме програмне забезпечення більше орієнтоване на облік та планування догляду за посівними культурами, що відкидає можливість прогнозування зростання та врожайності ряду рослин.

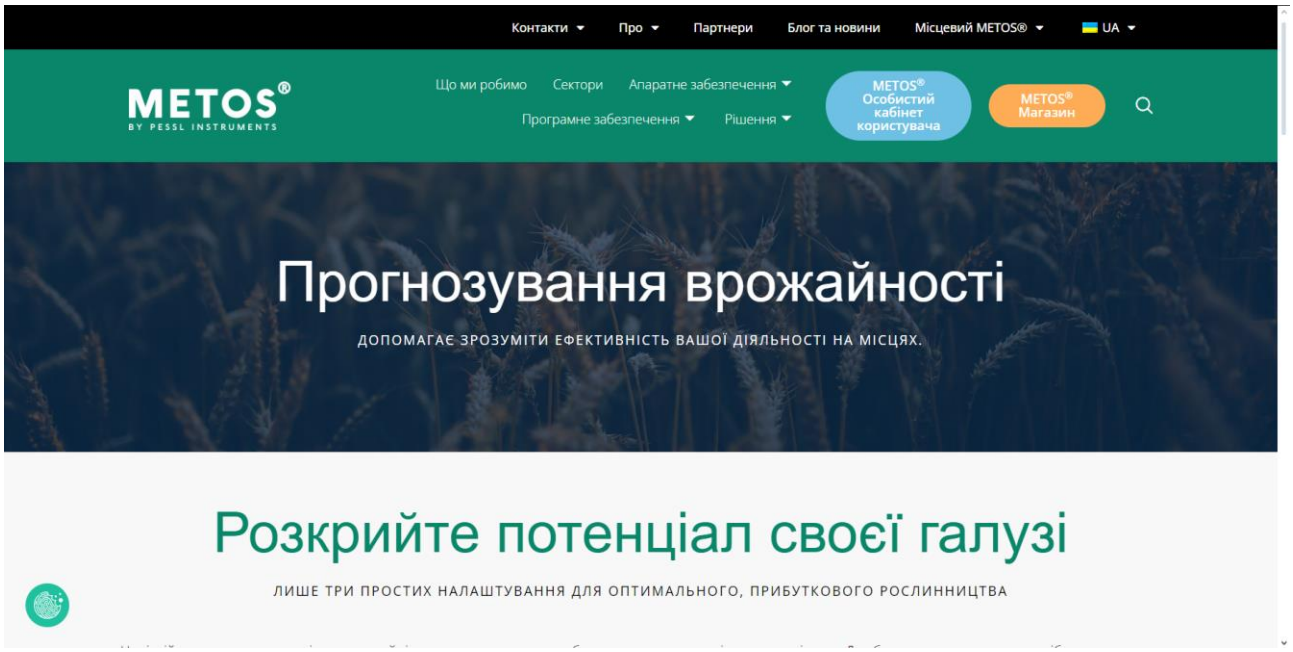


Рисунок 1.3 – Головна сторінка сайту «Metos»

З аналізу моделей-аналогів отримуємо:

- CISS GROUP: простий інтерфейс, але обмеженість у можливостях прогнозу, низька зручність та надійність;
- EOS: високотехнологічна система з високою якістю прогнозів, але є суттєві обмеження в часі;
- Metos: орієнтований на загальну прибутковість товару та моніторинг з плануванням власних ділянок, має обмежений список прогнозованих культур та ряд налаштувань.

Проведений аналіз показує, що існуючі методи мають свої переваги та недоліки. Традиційні підходи, засновані на агрономічних знаннях та статистичних даних, часто обмежені в точності та надійності. Розвиток сучасних технологій та доступність великих обсягів даних відкривають нові можливості для вдосконалення прогнозування. У цьому контексті використання методів машинного навчання, зокрема, методу випадкових лісів, може виявитися перспективним напрямком для підвищення точності та надійності прогнозів урожайності.

1.3 Аналіз технології обробки інформації в об'єкті дослідження

Аграрний сектор України має велику важливість для економіки та соціального розвитку країни. Зібрана інформація про врожайність культур стає ключовою для забезпечення продуктивності та планування господарської діяльності. Велика кількість даних про врожаї, які збираються у різних регіонах України, створює потужний інформаційний резерв для аналізу та прогнозування.

Сільське господарство – одна з галузей, де машинне навчання знайшло широке застосування. Воно дозволяє прогнозувати врожайність та визначати оптимальні стратегії для досягнення високих результатів. Аналіз даних з використанням алгоритмів машинного навчання допомагає сільськогосподарським підприємствам ухвалювати обґрунтовані рішення та зменшувати ризики. Такий підхід стає ключовим для забезпечення продуктивності та сталості сільського господарства.

Джерелами даних для аналізу врожайності культур є:

- державні статистичні органи, такі як Державна служба статистики України, що збирає та публікує статистику щодо вирощування культур у різних регіонах країни;
- сільськогосподарські підприємства та фермерські господарства, саме дані, які надходять від сільськогосподарських підприємств та фермерських господарств, містять інформацію про реальну врожайність та технології вирощування культур;
- спеціалізовані додатки та сенсори, в сучасний час використовуються різні сучасні технології для збору даних, включаючи додатки для моніторингу та сенсори для вимірювання показників ґрунту та атмосферних умов.

Дані про врожайність та агрономічні показники підлягають попередній обробці перед їх подальшим аналізом. Очищення, інтеграція та підготовка даних є важливим етапом для забезпечення якості та надійності аналізу. Використання сучасних інструментів та програмного забезпечення для обробки даних дозволяє враховувати всі необхідні параметри та виокремлювати важливі залежності.

Об'єктом дослідження є врожайність культур в Україні. Важливо враховувати специфіку географічних та кліматичних умов різних регіонів країни, а також вплив сільськогосподарських практик на результати вирощування культур. Аналіз цих факторів допомагає розробити моделі та методи прогнозування, які були б адаптовані до реалій аграрної галузі України.

Аналіз та обробка даних здійснюються за допомогою сучасних методів машинного навчання, включаючи випадкові ліси, дерева рішень, а також структурний аналіз і моделювання реляційних баз даних. Ці методи дозволяють створювати надійні прогнози та розробляти рекомендації для сільськогосподарських підприємств та інших учасників аграрного сектору.

У розділі проведено аналіз сучасного стану розвитку технологій машинного навчання та технології обробки інформації в об'єкті дослідження, зокрема в аграрному секторі України. Виявлено, що машинне навчання набуло значної популярності в різних сферах, включаючи сільське господарство.

Зібрана інформація та даний аналіз специфіки об'єкта дослідження вказують на важливість використання сучасних методів машинного навчання та обробки даних для прогнозування врожайності культур в Україні. Аналіз географічних та кліматичних умов, а також сільськогосподарських практик дозволяє розробити моделі, які враховують реалії аграрного сектору країни.

Застосування методів машинного навчання, таких як випадкові ліси та дерева рішень, разом із структурним аналізом і моделюванням реляційних баз даних, надає можливість створити надійні прогнози та рекомендації для сільськогосподарських підприємств, страхових компаній та інших учасників аграрного сектору. Результати досліджень можуть бути використані для планування та управління сільськогосподарською діяльністю, що важливо для стабільного розвитку аграрної галузі в Україні.

2 ПОСТАНОВКА ЗАДАЧІ

2.1 Постановка задачі на дослідження

Основною метою нашого дослідження є детальне вивчення методів машинного навчання та можливість їх використання у процесах прогнозування. В якості прикладу ми використовуємо інформаційну систему прогнозування врожайності культур. Ми акцентуємо увагу на аспекті машинного навчання та його застосуванні для покращення процесу прогнозування врожайності в сільському господарстві. Дослідження методів прогнозування нададуть можливість аграрному сектору та іншим зацікавленим сторонам в Україні ефективно вирішувати практичні завдання.

Для виконання поставленого завдання необхідно ретельно проаналізувати предметну область та розробити програмне забезпечення для прогнозування врожайності сільськогосподарських культур на території України. Воно включає в себе створення програмного засобу для обробки та аналізу великих обсягів даних щодо врожайності культур за попередні роки.

Завдання також включає роботу з великим обсягом даних, їхню очистку, підготовку для подальшого аналізу та використання у моделях машинного навчання. Програмний продукт повинен включати інтерфейс для зручного користування, який дозволить проводити аналіз та отримувати прогнози врожайності на основі введених даних.

Ключові завдання включають в себе створення програмного рішення для прогнозування врожайності культур на території України з використанням методів машинного навчання. Надзвичайно важливо враховувати аграрну специфіку України та надавати корисну інформацію для підтримки прийняття рішень в аграрному секторі та для досягнення максимального врожаю.

Основні завдання включають:

- аналіз предметної області, з визначенням архітектури побудови, з функціональним моделюванням та визначенням функціональних вимог;

моделювання потоків даних; вимоги до складу інформації (сутностей, атрибутів сутностей), що зберігатиметься в базі даних;

- вивчення аграрної специфіки України та врахування її у розробці моделей прогнозування;
- розробку та уточнення вимог до інтерфейсу, з використанням методології розробки програмного забезпечення;
- використання методів машинного навчання для точних та достовірних прогнозів врожайності;
- аналіз результатів прогнозування за обраним методом машинного навчання порівняно з реальними даними;
- визначення рекомендацій для підвищення якості прогнозів.

Ці завдання передбачають використання сучасних підходів до аналізу та обробки даних, а також розробку надійних і точних алгоритмів для прогнозування врожайності сільськогосподарських культур на Україні.

Для виконання необхідні: Операційна система Windows XP або вище, програмне забезпечення: CASE-засіб All Fusion Data Modeler (ERWin), All Fusion Process Modeler (BPWin), IDE PyCharm.

2.2 Визначення вхідних даних для прогнозу

Чітке визначення та докладний аналіз кожного параметра вхідних даних створюють науковий фундамент для розробки моделей прогнозування, спрямованих на досягнення високої точності та ефективності в прогнозуванні врожайності сільськогосподарських культур. Визначення параметрів є одним з ключових етапів процесі використання машинного навчання [9], воно має безпосередньо вплив на:

1. створення навчального датасету, щоб модель "навчилася" взаємозв'язкам між цими факторами та врожайністю для точного прогнозування;

2. створення тестового датасету, тобто дані за певний період використовуються для створення тестового датасету, який буде використаний для перевірки точності прогнозу;

3. аналіз тенденцій, тобто в моделі виявляють ключові фактори, які впливають на урожайність, зокрема, ті, які можуть бути взяті з минулих даних;

4. моделювання та оцінка, в котрих навчальні дані використовуються для тренування моделей, а тестові дані служать для оцінки їхньої точності та ефективності.

5. вдосконалення моделі, з зростанням обсягу даних про урожайність моделі можуть постійно оновлюватися та підтримуватися актуальними.

Визначення факторів, які мають вирішальне значення для урожайності сільськогосподарських культур, є критичним завданням для подальшого прогнозування та планування в аграрному секторі. Ми розглядаємо ключові фактори [10], які впливають на результативність врожайності та їх роль у визначенні майбутньої системи прогнозування.

Урожайність сільськогосподарських культур є результатом комплексного впливу численних факторів. Важливо враховувати наступні фактори:

1. погодні умови: погода має безперечний вплив на врожайність; температура, опади, вологість, інтенсивність сонячного випромінювання і вітер – всі ці параметри можуть визначити успіх чи невдачу врожаю; агрономи повинні вміти адаптувати агротехнології до конкретних погодних умов;

2. агротехнології: вибір правильних агротехнологій, включаючи обробку ґрунту, внесення добрив, і сівбу, грає вирішальну роль у врожайності;

3. ґрунт і його властивості: здоровий ґрунт, багатий на поживні речовини, забезпечує кращий ріст рослин; існує необхідність вивчати властивості ґрунту та розробляти стратегії його покращення;

4. рослинні матеріали: вибір сортів рослин для конкретного ґрунту та погодних умов може вирішити питання врожайності; необхідно підбирати оптимальні сорти залежно від обставин;

5. добрива та захист рослин: оптимальне внесення добрив і вчасне застосування засобів захисту рослин гарантують здорові рослини та врожайність;

6. планування і рішення: вибір розміщення полів, терміни сівби та збирання, а також вирішення проблем, таких як шкідники та хвороби, вимагають високої експертизи професіоналів.

Під час процесу прогнозування врожайності сільськогосподарських культур існує два основних підходи до аналізу та моделювання. Перший підхід передбачає врахування всіх можливих внутрішніх факторів, таких як вибір удобрень, агротехніка, планування полів та інші аспекти агрономічних дій. Він дозволяє враховувати всі можливі нюанси та агротехнічні аспекти, що можуть вплинути на урожайність культур.

Проте ця робота спрямована на інший підхід, який сконцентрований на визначенні та використанні зовнішніх факторів, які також мають велике значення для врожайності, однак на котрі людина зазвичай не може вплинути. Ці зовнішні фактори включають в себе кліматичні умови, географічне розташування та погодні умови, тип ґрунту та підкріплюється даними врожаїв минулих років.

Цей підхід дозволяє нам створювати моделі прогнозування, які можуть бути корисні для сільськогосподарського сектору та інших зацікавлених сторін, оскільки враховуються зовнішні фактори, які залишаються поза контролем фермерів та агрономів. Дані зовнішні фактори можуть бути рішучими для визначення врожайності та впливають на прийняття стратегічних рішень в аграрній сфері, надаючи можливість прийняти випереджаючі рішення за для згладжування можливих економічних та продовольчих криз.

Такий підхід дозволяє нам розширити область досліджень та прогнозів у сільському господарстві, зосереджуючись на тих факторах, які можуть бути важливими для українських фермерів та агрономів.

2.3 Збір даних

Для досягнення цих важливих цілей, необхідно визначити та ретельно розглянути види вхідних даних та їх джерела. Використання даних з минулих років у машинному навчанні є необхідним етапом для створення точних та надійних моделей прогнозування урожайності. Цей підхід дозволяє врахувати різноманітні фактори, що впливають на врожай, та забезпечує підґрунтя для подальшого вдосконалення алгоритмів прогнозу.

Аналіз історичних врожаїв є важливим етапом, де врахування попередніх врожаїв сприяє визначенню тенденцій та можливих ризиків. Для цього достатньо, в якості основного ресурсу, використати офіційний сайт державної служби статистики України [11] (рисунок 2.1) . Ми отримаємо такі данні: обсяг, вид вирощуваної культури та кількість врожаю відповідно до області.

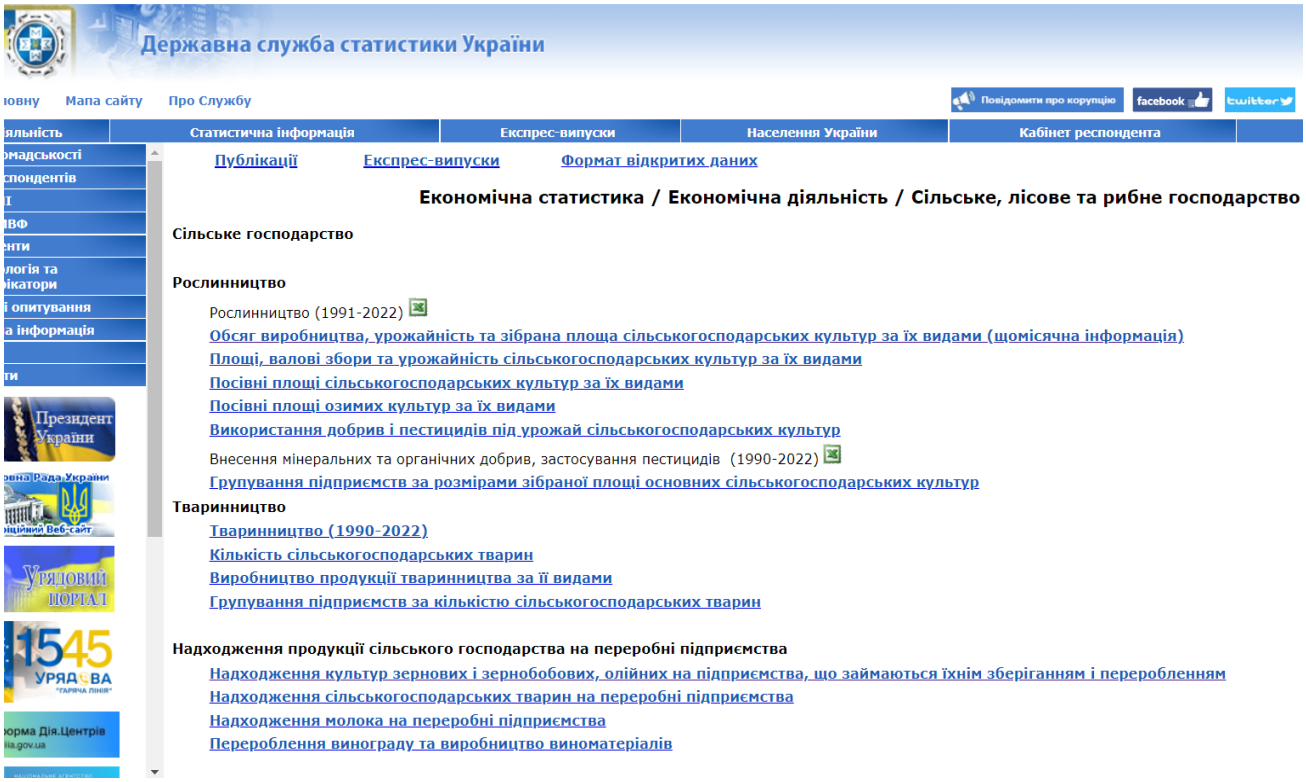


Рисунок 2.1 – Статистика сільського господарства

Поєднання географічних та агротехнічних параметрів, такі як розташування та тип ґрунту відносно нього , надають можливість узагальненого погляду на умови розвитку сільськогосподарських культур. Вони є важливими у визначенні можливостей вирощування певних видів, що сприяють ефективному використанню ресурсів.

Різноманіття ґрунтових властивостей [12] впливає на здатність ґрунту утримувати вологу, доступність поживних речовин для рослин та їхню стійкість до різних погодних умов. Типізація різновидів ґрунту надає можливість більш узагальнено зрозуміти основні його основні властивості такі як, текстура та хімічний склад, тобто кислотність.

Текстура ґрунту, така як крупнозернистий або дрібнозернистий, впливає на дренаж та здатність утримувати вологу. Це фактор, що може визначати ступінь вразливості до посухи та забезпечити важливу інформацію для прогнозування врожайності. Також надає змогу розуміти його властивості щодо утримання тепла. Наприклад, температурний режим чорноземів відрізняється від піщаних ґрунтів, що може впливати на час сходів та дозрівання рослин.

Кислотність ґрунту також важлива, оскільки вона впливає на доступність поживних речовин для рослин. Оптимальний рівень кислотності сприяє найбільш ефективному використанню добрив та покращує умови для зростання рослин.

Визначення типу ґрунту є одним з ключових елементів у прогнозуванні врожайності, оскільки це надає підґрунтя для аналізу, моделювання та розробки прогностичних стратегій, що враховують аспекти впливу ґрунтового середовища на рослинництво. Важливим буде ввести інформацію щодо основних типів ґрунту відповідно до областей спостереження, вона знаходиться у відкритому доступі у вигляді карт з відповідною номенклатурою [13] (рисунок 2.3).

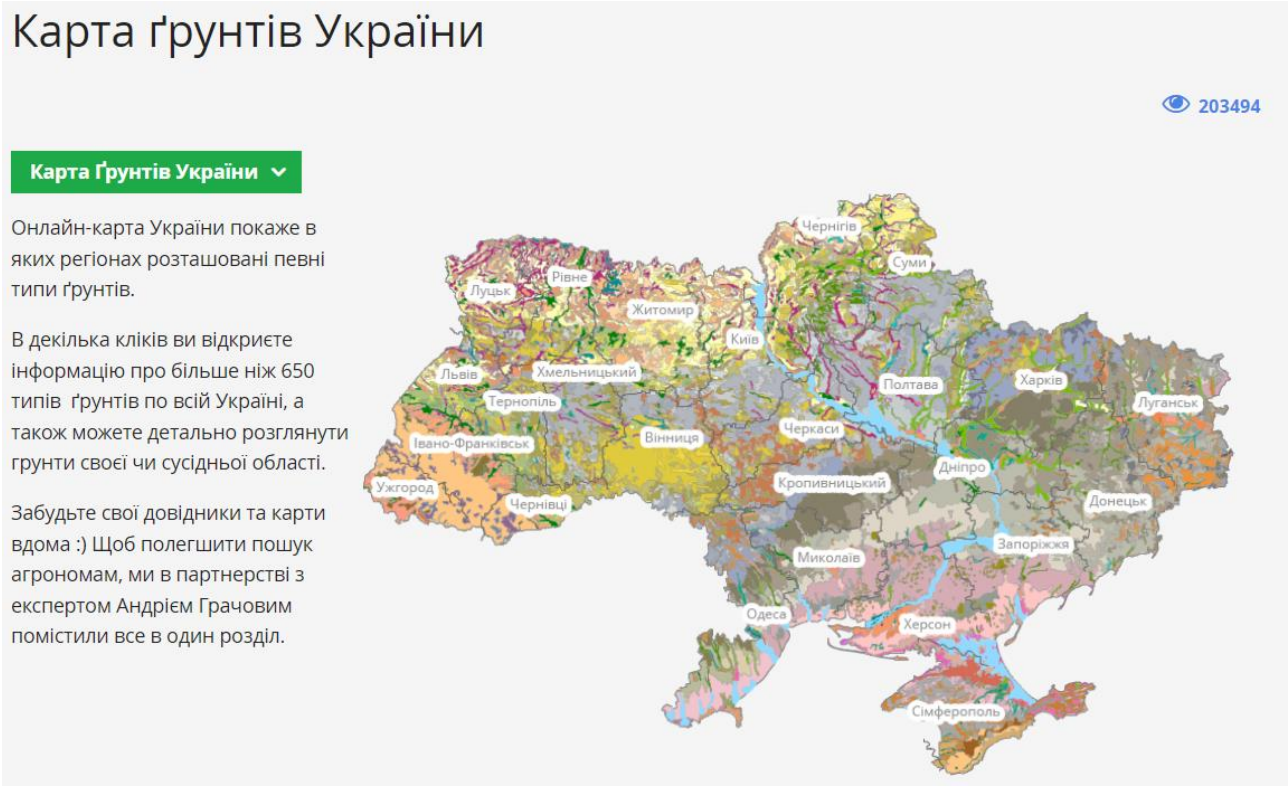


Рисунок 2.2 – Карта ґрунтів України

Однією з визначальних складових точного та надійного прогнозу урожайності є збір та використання погодних даних. Погода впливає на ріст рослин, їх розвиток та загальний врожай. З цього погляду, збір та аналіз погодних параметрів стає необхідним етапом в розробці прогнозуючих моделей.

Одним з ключових факторів є температура, яка впливає на фотосинтез та розподіл поживних речовин у рослинах. Висока температура може сприяти або, навпаки, заважати процесам зростання культур. Оптимальний розподіл опадів та їх рівень також важливі для забезпечення відповідних умов для розвитку рослин.

Збір даних про інтенсивність вітру, вологість повітря та інші погодні аспекти є також ключовим для розуміння мікрокліматичних умов. Ці дані враховуються в прогнозних моделях для створення більш точних і адаптивних прогнозів. Окрім того, вони дозволяють аграріям та дослідникам аналізувати вплив погодних аномалій на врожай та вживати заходи для зменшення ризиків.

Усі ці фактори підкреслюють важливість систематичного та точного збору погодних даних для побудови надійних прогнозів урожайності. Тому за для

створення відповідних датасетів варто звернутися до архівів щоденників погоди [14]. Більш за все, варто приділяти значення екстремумам, та враховувати середнє значення показників за місяць.

Рисунок 2.3 – Архів погоди meteoblue

2.4 Підготовка даних, визначення цільового вектору

В цілому, процес підготовки даних для прогнозування з використанням машинного навчання включає кілька ключових етапів, спрямованих на оптимізацію якості та ефективності прогнозів урожайності (деякі з них вже було зазначено у попередньому підрозділі):

1. збір та отримання даних: початковий етап передбачає систематичний збір різноманітних даних, необхідних для ефективного прогнозу врожайності, це включає в себе історичні дані про врожаї, кліматичні умови, агротехнічні

параметри, властивості ґрунту та інші фактори, які можуть впливати на вирощування сільськогосподарських культур;

2. очищення та фільтрація даних: перед використанням даних для навчання моделі важливо провести їх очищення та фільтрацію, цей крок допомагає виявити та виправити можливі аномалії чи неточності, які можуть впливати на якість прогнозу;

3. нормалізація та стандартизація даних, що допомагають підготувати їх для ефективного навчання моделі, хоча існують методи, що не чутливі до масштабу, ці процеси можуть поліпшити швидкість навчання;

4. виділення ознак: важливим аспектом є відбір ключових ознак для використання в моделі, існує ряд методів, що можуть автоматично оцінювати їх важливість, але важливо ретельно підготувати їх набір перед використанням;

5. розділення даних: для ефективної оцінки ефективності моделі необхідно розділити дані на тренувальний та тестовий набори, це дозволяє перевірити, наскільки добре модель справляється з новими даними;

6. обробка відповідей (цільових показників): задача визначення цільових показників залишається актуальною, це може включати в себе створення цільового вектору, представлення значення, яке модель повинна прогнозувати.

Цільовий вектор [15] в машинному навчанні є частиною даних, яку ми намагаємося прогнозувати або передбачити. У випадку прогнозування урожайності за допомогою методу випадкових лісів, цільовий вектор містить інформацію про очікувані врожаї або вихідні дані, які ми спробуємо передбачити за допомогою моделі.

Наприклад, якщо ми розглядаємо прогноз врожайності пшениці, цільовий вектор може містити кількість врожаю пшениці на певний період часу. Мета полягає в тому, щоб модель на основі інших вхідних даних, таких як кліматичні умови, агротехнічні параметри тощо, навчилася ефективно передбачати цей цільовий вектор.

Отже, цільовий вектор представляє собою ціль, до якої спрямована прогностична модель, і яку вона намагається точно передбачити на основі

доступних даних. Тобто вихідні данні нашої практичної частини це кількісна міра можливого врожаю, що у нашому випадку співпадає з визначенням цільового вектору.

На основі опису вхідних та вихідних даних можна представити функціональну модель майбутньої системи відповідно до стандарту IDEF0, що показані на рисунках 2.4 та 2.5.

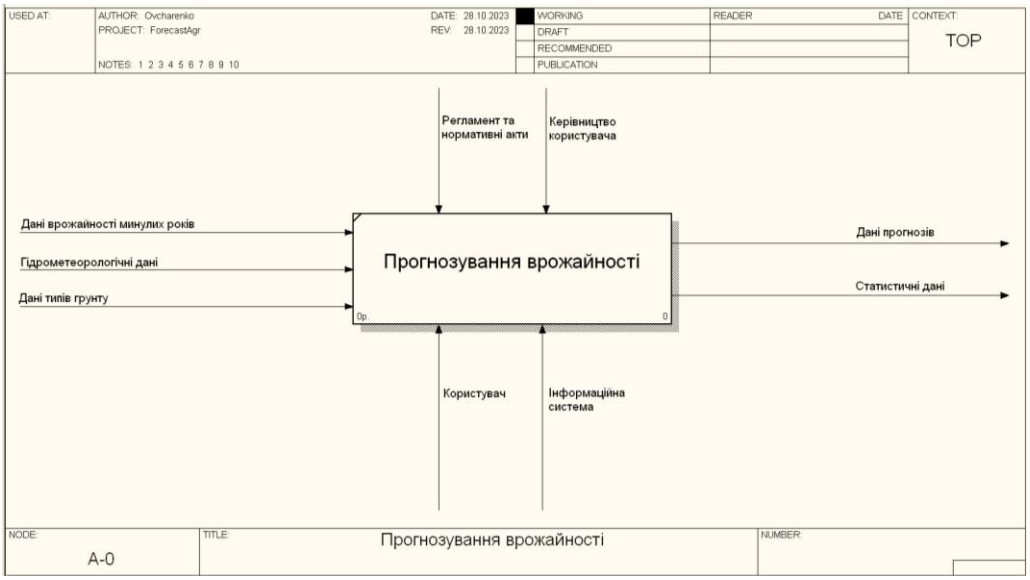


Рисунок 2.4 - Функціональна модель за стандартом IDEF0

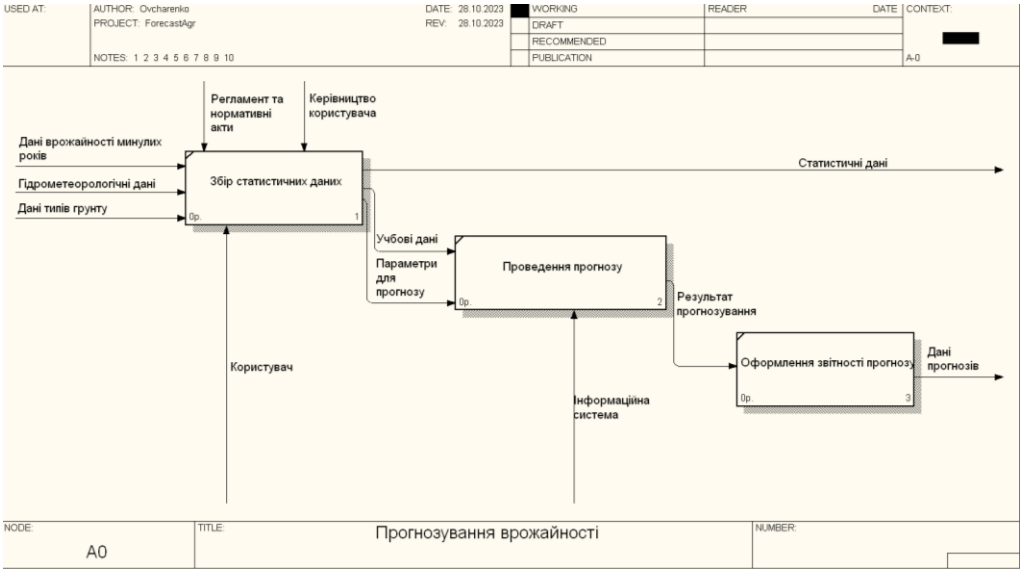


Рисунок 2.5 – Декомпозиція першого рівня

3 ДОСЛІДЖЕННЯ МЕТОДІВ ВИРІШЕННЯ ЗАДАЧІ

Розділ має на меті ретельне вивчення різних підходів у сфері машинного навчання, зокрема в контексті прогнозування урожайності.

Навчання [9], подібно до інтелекту, охоплює такий широкий спектр процесів, що його важко точно визначити. Словникове визначення включає фрази типу "отримувати знання, розуміння або навички шляхом вивчення, інструкцій або досвіду" та "модифікація поведінкової тенденції досвідом". Щодо машин, можна сказати дуже широко, що машина вчиться, кожного разу, коли вона змінює свою структуру, програму або дані (на основі своїх вхідних даних або відповідно до зовнішньої інформації) таким чином, що її очікувана майбутня продуктивність поліпшується. Деякі з цих змін, такі як додавання запису до бази даних, зручно вписуються в інші дисципліни і не обов'язково краще зрозумілі під назвою "навчання". Але, наприклад, коли продуктивність машини розпізнавання мови поліпшується після прослуховування кількох зразків мови людини, ми вважаємо, що в цьому випадку машина вивчила.

Слід зазначити, машинне навчання [9] – це обширна галузь, що охоплює різні методи, серед яких глибоке навчання що, досягло рівня уваги громадськості та інвестицій промисловості, якого ніколи раніше не було в історії штучного інтелекту, є лише одним з напрямів.

Проте це не єдиний успішний вид машинного навчання. Можна заявити, що більшість алгоритмів машинного навчання, які використовуються в промисловості сьогодні, не є алгоритмами глибокого навчання. Глибоке навчання [16] не завжди є найкращим інструментом для завдання — іноді не вистачає даних для застосування глибокого навчання, а іноді проблему краще вирішити іншим алгоритмом.

У цьому розділі має на меті ретельне вивчення різних підходів у сфері машинного навчання, зокрема в контексті прогнозування урожайності. Розглядаючи ці різноманітні підходи, ми можемо краще розуміти їх сильні та

слабкі сторони, а також визначити, чому вибір певного методу, наприклад, методу випадкових лісів, є обґрунтованим для конкретного випадку прогнозування урожайності.

3.1 Порівняння методів машинного навчання

Метод часових рядів [17] – це підхід до прогнозування, призначений для аналізу та передбачення змін часових рядів. Цей метод розглядає дані як послідовність точок у часі та використовує попередні значення для прогнозування майбутніх.

Алгоритм часових рядів в машинному навчанні використовується для аналізу та прогнозування динаміки даних у вигляді послідовностей, де час грає важливу роль. Зазвичай використовуються моделі, такі як ARIMA [18] (авторегресійна інтегрована змінна середніх) або моделі на основі нейронних мереж. Ми розглянемо більш детально перший випадок:

Цей алгоритм (рисунок 3.1) надає структурований підхід до роботи з часовими рядами у машинному навчанні, дозволяючи ефективно аналізувати та прогнозувати динаміку часових даних.

ARIMA (Autoregressive Integrated Moving Average) – це статистичний метод аналізу та прогнозу часових рядів. Цей метод включає авторегресію (AR), інтеграцію (I) та ковзне середнє (MA) [19].

Авторегресія (AR) визначає, як поточне значення часового ряду залежить від попередніх значень у самому ряді (3.1).

$$\mathrm{Xt} = \phi_1 \cdot \mathrm{X}_{t-1} + \phi_2 \cdot \mathrm{X}_{t-2} + \cdots + \phi_p \cdot \mathrm{X}_{t-p} + \epsilon_t \qquad (3.1)$$

Параметр p вказує на кількість попередніх значень, що використовуються для прогнозу, ϕ_1 , ϕ_2 ,…, ϕ_p - параметри авторегресії, ϵ_t – біла шумова помилка.

Інтеграція (I) вказує на необхідність диференціювання часового ряду для становлення його стаціонарним (вільним від тренду та сезонності), представлено формулою 3.2.

$$Y_t = X_t - X_{t-1} \tag{3.2}$$

Параметр d вказує на кількість диференціювань, необхідних для досягнення стаціонарності, у нашому випадку формула представлена для d=1, а Y_t – стаціонарний часовий ряд.

Ковзне середнє (MA) представляє собою модель ковзного середнього, що вказує на залежність поточного значення ряду від попередніх значень шумового члену, відображено формулою 3.3.

$$X_t = \theta_1 + \epsilon_{t-1} + \theta_2 \cdot \epsilon_{t-2} + \cdots + \theta_q \cdot \epsilon_{t-q} + \eta_t \tag{3.3}$$

Параметр q вказує на кількість попередніх шумових членів, що використовуються для прогнозу, де θ - параметри ковзного середнього, η_t - біла шумова помилка.

Процес побудови ARIMA-моделі:

1. перевірка стаціонарності, визначення необхідності диференціювання для становлення ряду стаціонарним;
2. визначення параметрів (p, d, q), тобто використання аналізу автокореляції та часткової автокореляції для визначення p та q, визначення d на основі перевірки стаціонарності;
3. застосування підібраних параметрів для побудови ARIMA-моделі;
4. навчання моделі за допомогою тренувального набору даних;
5. оцінка ефективності моделі за допомогою тестового набору даних.
6. використання навченої моделі для прогнозування майбутніх значень часового ряду.

ARIMA дозволяє моделювати та прогнозувати ряди з урахуванням їхньої динаміки та елементів часу.

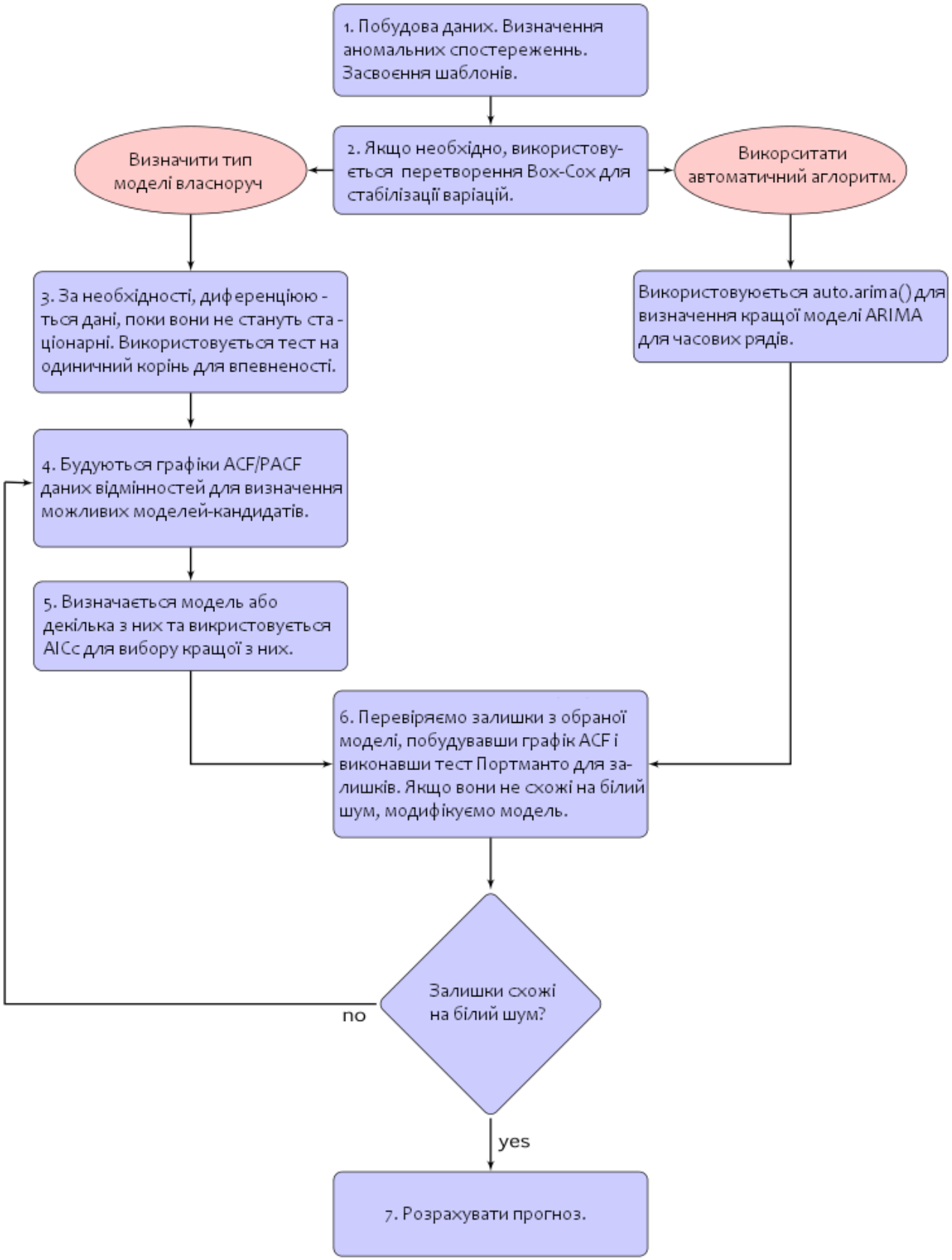


Рисунок 3.1 – Алгоритм ARIMA

Основні переваги та недоліки зазначено у таблиці 3.1.

Таблиця 3.1 – Метод часових рядів

Переваги	Недоліки
Врахування динаміки в часі: метод часових рядів дозволяє враховувати зміни в даних в залежності від часу, що допомагає ліпше виявляти тренди та сезонність.	Чутливість до випадкових збурень: метод може погано справлятися зі збуреннями або випадковими подіями, які важко передбачити.
Простота інтерпретації: оскільки цей метод базується на історичних даних та залежить від попередніх значень, результати прогнозування легше інтерпретувати.	Неефективність для динамічних змін: якщо в часовому ряді відбуваються динамічні зміни, такі як зміна тренду чи швидкості змін, метод може показати неприйнятні результати.
Ефективність для стабільних рядів: вони добре працюють для стабільних часових рядів, де основні характеристики лишаються майже незмінними з часом.	Обмеженість для врахування зовнішніх факторів: метод часових рядів не завжди ефективний для врахування зовнішніх факторів, які можуть впливати на часовий ряд.

Метод часових рядів є потужним інструментом для прогнозування, особливо в стабільних умовах. Однак, перед використанням цього методу, важливо ретельно аналізувати природу даних та їх мінливість для того, щоб визначити його придатність для конкретної задачі прогнозування.

Метод часових рядів може бути не найкращим вибором для прогнозування урожайності. Оскільки деякі значення, наприклад сама врожайність, можуть бути не стабільними, окрім цього, якщо урожайність супроводжується великими випадковими змінами, які важко передбачити або пояснити, метод часових рядів може виявитися чутливим до таких непередбачуваних коливань.

Лінійна регресія [20]— це метод машинного навчання, який моделює залежність між залежною змінною (в нашому випадку, врожайністю) та однією або більше незалежними змінними (вхідними параметрами), припускаючи, що ця залежність є лінійною. Метою є знаходження оптимальної лінії (лінії регресії), яка найкраще відображає відношення між змінними, схематично відображено на рисунку 3.2.

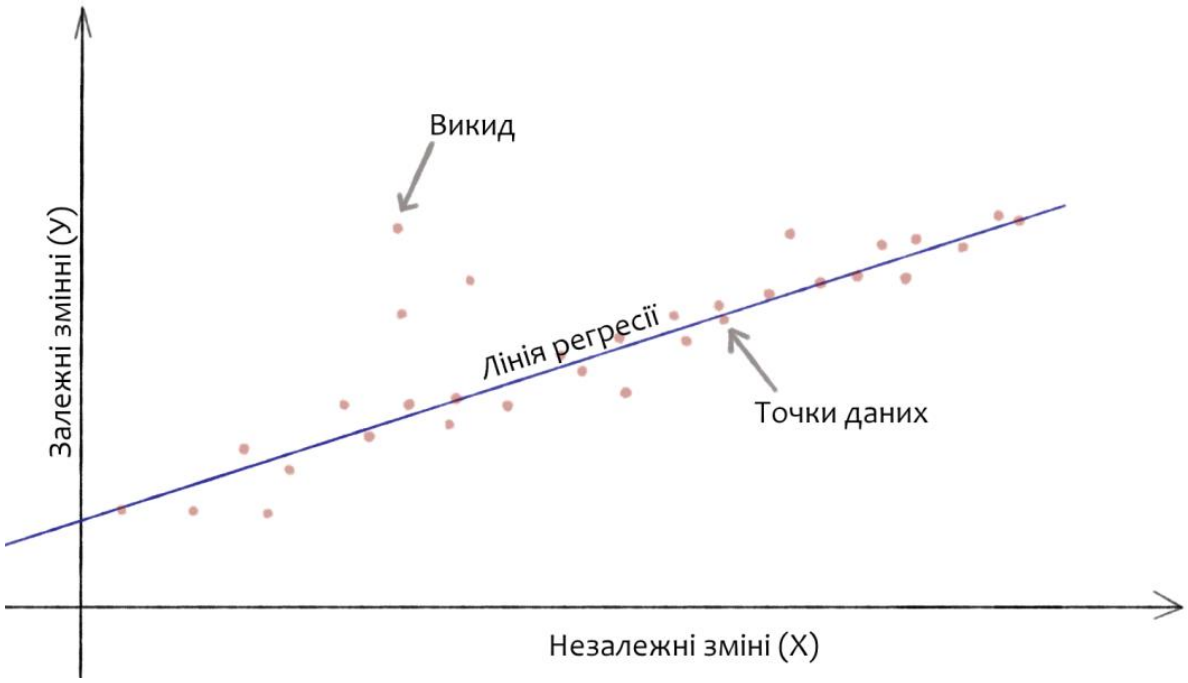


Рисунок 3.2 – Лінія регресії

Цей метод широко використовується у прогнозуванні та аналізі даних. Основна ідея полягає в тому, щоб знаходити лінійну функцію, яка найкращим чином відображає залежність між вхідними та вихідними даними.

Першим кроком підготовляємо дані: визначаємо вхідні ознаки (незалежні змінні) та вихідну змінну (залежну змінну). Наступним кроком є визначення функції лінійної регресії. Припускаємо, що залежність між вхідними та вихідною змінною є лінійною. Модель лінійної регресії вказана формулою 3.4.

$$y = b_0 + b_1 \cdot x_1 + b \cdot x_2 + \cdots + b_n \cdot x_n + \epsilon \quad (3.4)$$

Де y - вихідна змінна, $x_1, x_2, \ldots, x_n$ – вхідні ознаки, $b_1, b_2, \ldots, b_n$ – коефіцієнти регресії, є - помилка.

Далі необхідно провести визначення критерію функції втрат, який виміряє, наскільки добре модель відповідає даним. Зазвичай використовується середньоквадратична помилка, що відображено формулою 3.5.

$$MSE\ (\text{Mean Squared Error}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (3.5)$$

Де n - кількість спостережень, y_i - фактичне значення, $y^\wedge_i$ - прогнозоване значення.

Наступним кроком є оптимізація коефіцієнтів. Проводиться мінімізація функції втрат за допомогою методу найменших квадратів, тобто обчислення значень коефіцієнтів, що мінімізують суму квадратів різниць між фактичними та прогнозованими значеннями.

Отримані значення коефіцієнтів використовуються для оцінки відносин між вхідними та вихідною змінною. Використовуючи навчену модель, робимо прогнози для нових даних.

Лінійна регресія [21] – це простий та ефективний метод, який широко використовується у багатьох галузях для аналізу та прогнозування. Основні переваги та недоліки котрого зазначено у таблиці 3.2.

Таблиця 3.2 – Лінійна регресія

Переваги	Недоліки
Простота та зрозумілість: лінійна регресія є простим методом, легким для розуміння та інтерпретації результатів.	Лінійність: обмеженість в передбачуваності в тих випадках, коли відносини між змінними не лінійні.

Продовження таблиці 3.2

Переваги	Недоліки
Швидкодія: лінійна регресія може швидко навчатися і використовуватися для передбачення.	Чутливість до викидів: вразливість до аномальних значень у даних, що може впливати на якість моделі.
Ефективність для стабільних рядів: вони добре працюють для стабільних часових рядів, де основні характеристики лишаються майже незмінними з часом.	Наявність взаємовідносин між змінними: якщо взаємодія між змінними складна або нелінійна, лінійна регресія може не врахувати цю складність.

Однак знову, скоріш за все не найкращий вибір, бо деякі взаємодії між різними факторами є дещо складніші, ніж є на перший погляд, наявна вірогідність нелінійних відносин, через що метод може недоцільно моделювати ці взаємодії.

Метод опорних векторів (SVM або Support Vector Machines*)* [22] є алгоритмом машинного навчання, який використовується як для класифікації, так і для регресії. Цей метод старається знайти оптимальну гіперплощину, яка найкраще розділяє дані у просторі так, щоб максимізувати розрив між класами, однак цей метод також можна використовувати так і для прогнозу значення регресії.

Розглянемо основні концепції методу опорних векторів [23]. Гіперплощина - це (n-1)-розмірна площина, яка розділяє n-розмірний простір на два класи. У двовимірному просторі гіперплощина - це лінія, у тривимірному - це площина, і так далі. Опорні вектори - це точки даних, які лежать найближче до гіперплощини, вони визначають положення та орієнтацію гіперплощини.

На рисунку 3.3 розглянуто приклад гіперплощини у тривимірному просторі, що розподіляє точки даних з різними ознаками, що для комфорту зазначено різними кольорами.

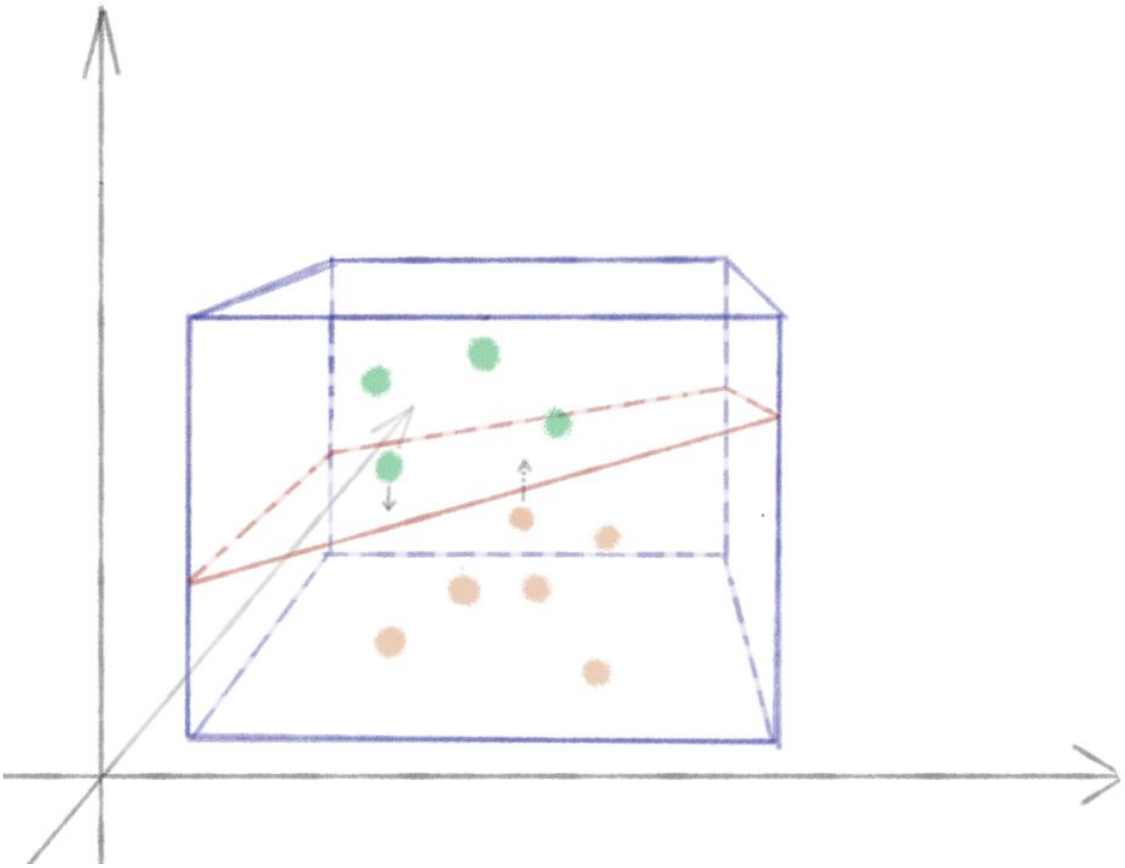


Рисунок 3.3 – Приклад відділяючої площини

Розподільча функція визначає відстань між точкою та гіперплощиною. Розподільчі проміжки - це зони навколо гіперплощини, що визначають класифікацію.

Ядро - це функція, яка визначає внутрішній продукт між двома точками в просторі вищого розміру. Використання ядер дозволяє вирішувати нелінійні задачі класифікації та регресії.

Етапи роботи методу опорних векторів:

- визначення простору ознак, в якому розташовані дані;
- визначення векторів ознак, які описують дані;
- визначення, чи вирішуємо завдання класифікації, де розділяємо дані на класи, чи регресії, де прогнозуємо числове значення;
- вибір ядра для вирішення задачі;
- вибір параметрів ядра, таких як ступінь полінома, які впливають на форму гіперплощини;

- максимізація відстані між гіперплощиною та опорними векторами;
- вирішення оптимізаційної задачі за допомогою методу квадратичного програмування [23];
- використання оптимальної гіперплощини для класифікації нових даних чи прогнозування числових значень.

Метод опорних векторів є потужним інструментом для вирішення різноманітних задач машинного навчання та має широкий спектр застосувань у великій кількості галузей. Основні переваги та недоліки зазначено у таблиці 3.3.

Таблиця 3.3 – Метод опорних векторів

Переваги	Недоліки
Ефективність в просторах великої розмірності: SVM ефективно працює в просторах великої розмірності, таких як задачі з багатьма ознаками.	Чутливість до великого обсягу даних: Для великих обсягів даних модель SVM може вимагати значних обчислювальних ресурсів.
Ефективність в умовах обмеженої кількості даних: Особливо ефективний, коли кількість зразків обмежена.	Важкоінтерпретованість: Отримана гиперплощина може бути важко інтерпретована у випадку високих розмірностей.

Основному призначенню методу більше відповідає класифікація, однак все ще можливо вести прогнози через регресію, можна запримітити дещо спільне з раніше згаданою лінійною регресією, однак відмінністю можна виділити більшу стійкість до викидів, та можливість розподіляти за більшою кількістю ознак [24].

Суттєвих недоліків для роботи з поданим методом немає, однак особливих переваг він також не надає, однак враховуючи обсяги даних кліматичних змін можу зробити припущення, що знадобиться значний обчислювальний ресурс.

Метод випадкових лісів (Random Forests) [24, 27] є ансамблевим методом машинного навчання, що базується на ідеї агрегації рішень багатьох дерев рішень. Основна ідея полягає в тому, щоб навчити кожне дерево на випадковому підмножині тренувальних даних і за рахунок голосування або середнього значення рішень кожного дерева отримати кінцевий прогноз. Загальне уявлення алгоритму продемонстровано на рисунку 3.5. Математичну складову розглянемо у наступному підрозділі.

Дерево рішень [3] — це дерево, внутрішні вузли якого є тестами, а листові вузли — категоріями. Наведемо приклад на рисунок 3.4. Дерево рішень призначає номер класу (або вихід) вхідному шаблону шляхом фільтрації шаблону через тести в дереві. Кожен тест має взаємовиключні та вичерпні результати.

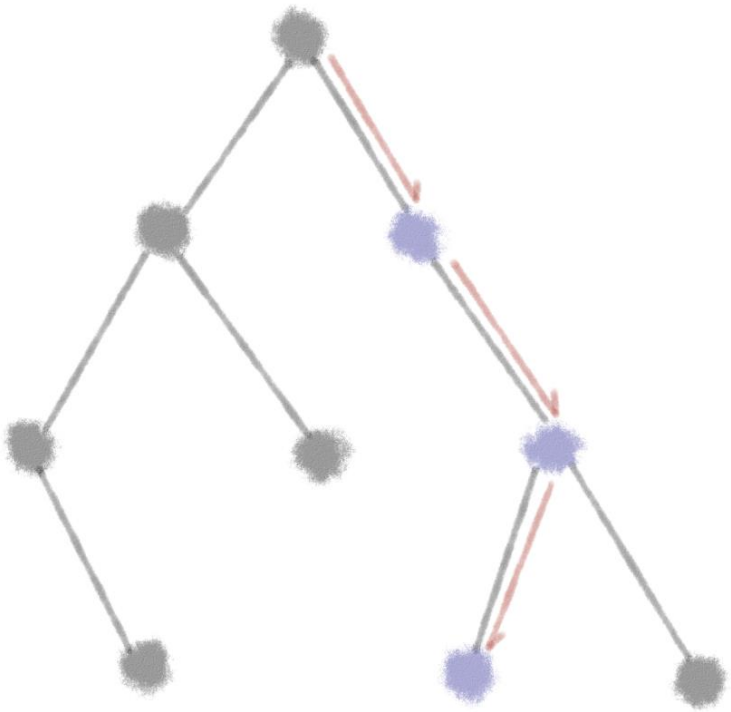


Рисунок 3.4 – Схематичне зображення дерева рішень

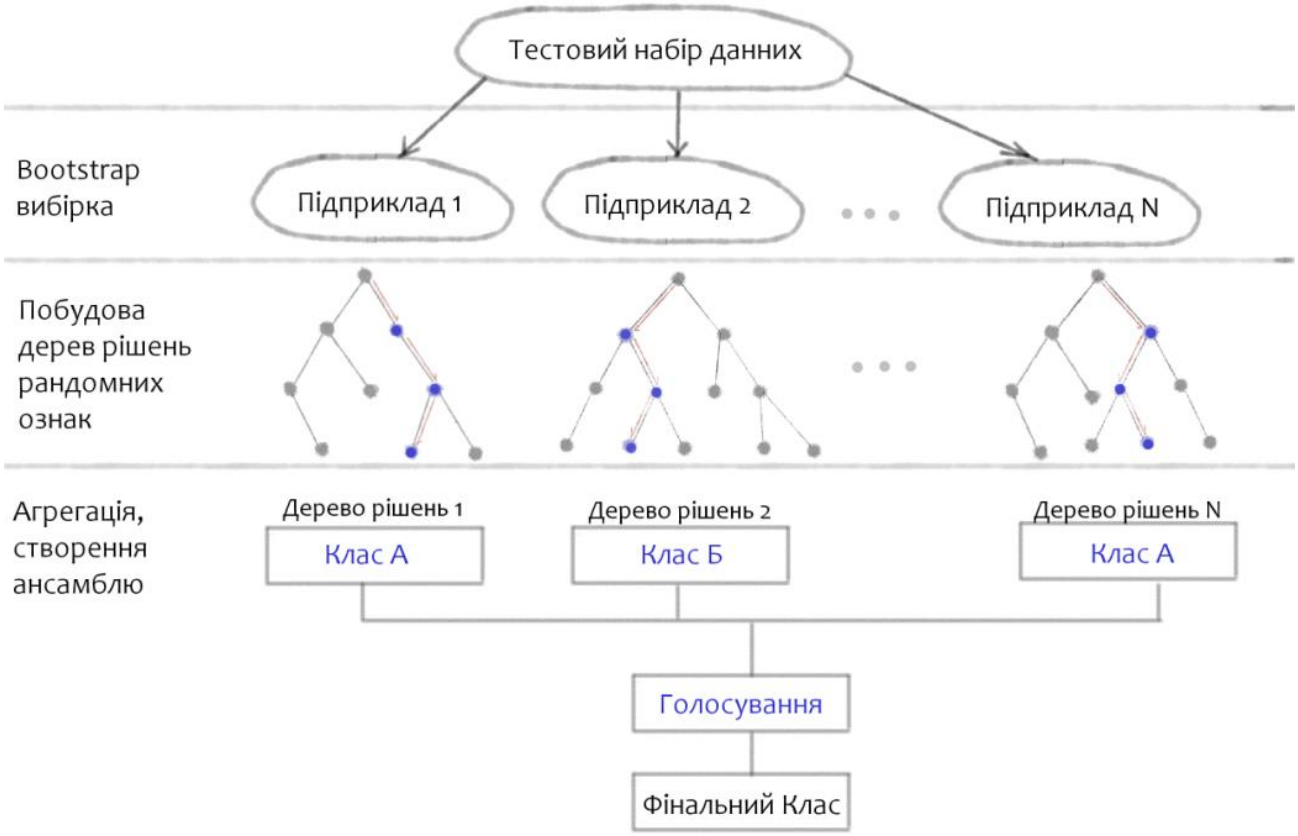


Рисунок 3.5 – Алгоритм навчання випадкових лісів

Метод використовує два види випадковості: випадковий вибір підмножини ознак для навчання кожного дерева і випадковий вибір прикладів для навчання кожного дерева. Це дозволяє зменшити переоцінку (overfitting) і зробити модель більш стійкою. Основні переваги та недоліки цього методу зазначено у таблиці 3.4.

Таблиця 3.4 – Метод випадкових лісів

Переваги	Недоліки
Висока точність: випадкові ліси зазвичай надають високу точність в прогнозуванні і класифікації, оскільки вони використовують ансамбль дерев рішень.	Велика обчислювальна складність: велика кількість дерев та їх навчання може вимагати значних обчислювальних ресурсів.

Продовження таблиці 3.4

Переваги	Недоліки
Взаємодія з великою кількістю змінних: можливість обробки великої кількості змінних без явного вибору ознак, що дозволяє враховувати багато різних факторів при прийнятті рішень.	Важко інтерпретується: однією з недоліків є складність інтерпретації моделі через велику кількість дерев та їх взаємодію.
Зменшення переоцінки: через випадковий вибір підмножини ознак для кожного дерева і випадкове вибір прикладів для навчання, можливо зменшення переоцінки (overfitting).	Потребує велику кількість даних: випадкові ліси можуть потребувати багато даних для досягнення оптимальної ефективності.
Виявлення важливості ознак: випадкові ліси можуть надати важливість кожній ознаці, що допомагає розуміти вплив різних факторів на прогноз.	
Стійкість до викидів: здатність адаптуватися до викидів або неправильних даних завдяки великій кількості дерев і випадковому вибору прикладів.	

Серйозним недоліком може стати обмеження обчислювальних ресурсів, велика кількість дерев може бути непрактичною, і інші методи можуть бути ефективнішими, однак метод добре працює з великими обсягами даних, здатний обробляти багато ознак, стійкий до перенавчання.

Переглянувши ряд методів отримуємо таку картину:

- при великих обсягах даних з багатьма ознаками і важлива точність, випадкові ліси можуть бути хорошим вибором;
- якщо завдання пов'язане з часовими залежностями і прогнозуванням, часові ряди можуть бути бажаними;
- якщо надана перевага простим і інтерпретованим моделям, лінійна регресія може бути відмінним варіантом;
- при складних даних з нелінійними залежностями, метод опорних векторів може бути корисним.

Зважуючи усі недоліки та переваги, для практичної частини я продовжу досліджувати *метод випадкових лісів*.

3.2 Математичний опис методу випадкових лісів

Нехай маємо набір даних $D=\{(X_1,Y_1),(X_2,Y_2),...,(X_N,Y_N)\}$, де X_i – вектор ознак, Y_i – відповіді, і N – кількість прикладів. Для застосування методу необхідно пройти такі кроки:

1. з кожним разом формуємо випадкову бутстреп-вибірку D_k розміром N з вихідного набору даних D, тобто ми вибираємо N прикладів з повтореннями;

2. будуємо дерева рішень, на кожному вузлі дерева обирається випадковий піднабір ознак (зазвичай, квадратний корінь від загальної кількості ознак) і область розбиття визначається за допомогою однієї з ознак цього піднабору; процес розбиття повторюється для кожного вузла досягнення певного критерію зупинки (наприклад, максимальна глибина дерева або мінімальна кількість прикладів у листовому вузлі);

3. повторюємо процес побудови дерева (процес побудови лісу) для кожної бутстреп-вибірки, у кінці отримуємо колекцію незалежних дерев (ліс);

4. для класифікації кожного нового прикладу проводимо голосування серед усіх дерев, однак, якщо ми вирішуємо задачу регресії, то можемо взяти середнє значення відповідей, отриманих від усіх дерев.

Це базовий алгоритм для методу випадкових лісів [2, 25, 26]. Метод випадкових лісів добре підходить для класифікації та регресії, а його випадковість у побудові дерев забезпечує стійкість до перенавчання та високу точність прогнозів.

3.3 Удосконалення методу випадкових лісів

Метод випадкових лісів, вже здатний до вражаючої точності, але може бути вдосконаленим в подальшому за допомогою ряду альтернативних підходів. Наведено кілька можливостей удосконалення, а також концентрація на модифікації з використанням ваг.

1. Ваги для прикладів: Додавання ваг для кожного прикладу дозволяє змінювати їхню важливість під час навчання моделі. Це особливо корисно, коли деякі приклади мають більший вплив або важливість для прогнозування врожаю. Додавання ваг дозволяє збалансувати вплив різних класів або підгруп у тренувальних даних.

2. Вибіркова підміна вибірки: Замість вибору всіх прикладів для навчання кожного дерева, можна розглядати вибіркову підміну, де лише частина прикладів використовується для кожного дерева. Це може поліпшити різноманіття моделі та зменшити кореляцію між деревами.

3. Композиція з іншими методами: Розглядаємо можливість комбінування методу випадкових лісів з іншими методами машинного навчання, такими як градієнтний бустінг або нейронні мережі, для підвищення ефективності та робустності моделі.

4. Використання інших оцінювачів важливості: Додавання альтернативних методів оцінювання важливості ознак, таких як permutation importance чи SHAP (Shapley Additive exPlanations), може збільшити точність оцінки важливості функцій.

5. Адаптація ваг для різних класів: Розглядаємо можливість індивідуального визначення ваг для різних класів в задачах класифікації для забезпечення більшого впливу важливих класів на модель.

Додавання ваги при розбитті вузла в дереві призведе до модифікації процедури випадкового лісу. Ось спрощений опис загального підходу:

1. Зміна Розбиття Вузла: При обранні поділу вузла для дерева на кожному етапі вибору розбиття додається вага для кожного випадкового вибору атрибуту. Вага може бути задана вручну або обчислюватися на основі якихось критеріїв.

2. Обчислення Ймовірностей: Після обрання ваги для кожного атрибута можна обчислити ймовірності вибору кожного атрибута для розбиття вузла. Зазвичай ймовірність обирають пропорційно вазі.

3. Модифікація Процесу Зростання Дерева: При розбитті вузла кожне дерево у лісі робить вибір атрибуту, враховуючи ймовірності, обчислені на попередньому кроці.

4. Збільшення Ваги Важливих Екземплярів: Якщо є ваги для екземплярів у наборі даних (наприклад, приблизно 30% екземплярів випадковим чином викидаються при кожному розбитті), можна також враховувати вагу при обчисленні функції втрат або інших критеріїв, які визначають якість розбиття.

Використовуючи ваги для прикладів, можна досягти більшої гнучкості та точності моделі в контексті прогнозування врожаю. Однак, важливо збалансувати ваги, щоб уникнути перенавчання чи недонавчання моделі. На основі конкретних даних, варто визначити які найбільш показові та впливові на загальний прогноз. Оцінка важливості прикладів може бути проведена на основі ряду факторів, що впливають на значущість прикладу дослідження. Ось деякі можливі критерії, які можна врахувати при визначенні ваг:

1. Історія врожаю: Приклади, які представляють собою історію врожаю протягом кількох сезонів, можуть мати більшу вагу, оскільки вони дозволяють врахувати динаміку вирощування культур протягом часу.

2. Екстремальні події: Важливість прикладів, пов'язаних з екстремальними погодними умовами, може бути підсилена, оскільки вони можуть суттєво впливати на врожай.

4. Ґрунтові характеристики: Приклади, що представляють різні грунтові умови, можуть мати важливість з точки зору адаптації культур до різних середовищ.

5. Результати попередніх прогнозів: Якщо є дані про попередні прогнози для тих самих прикладів, їх можна використовувати для визначення ваги.

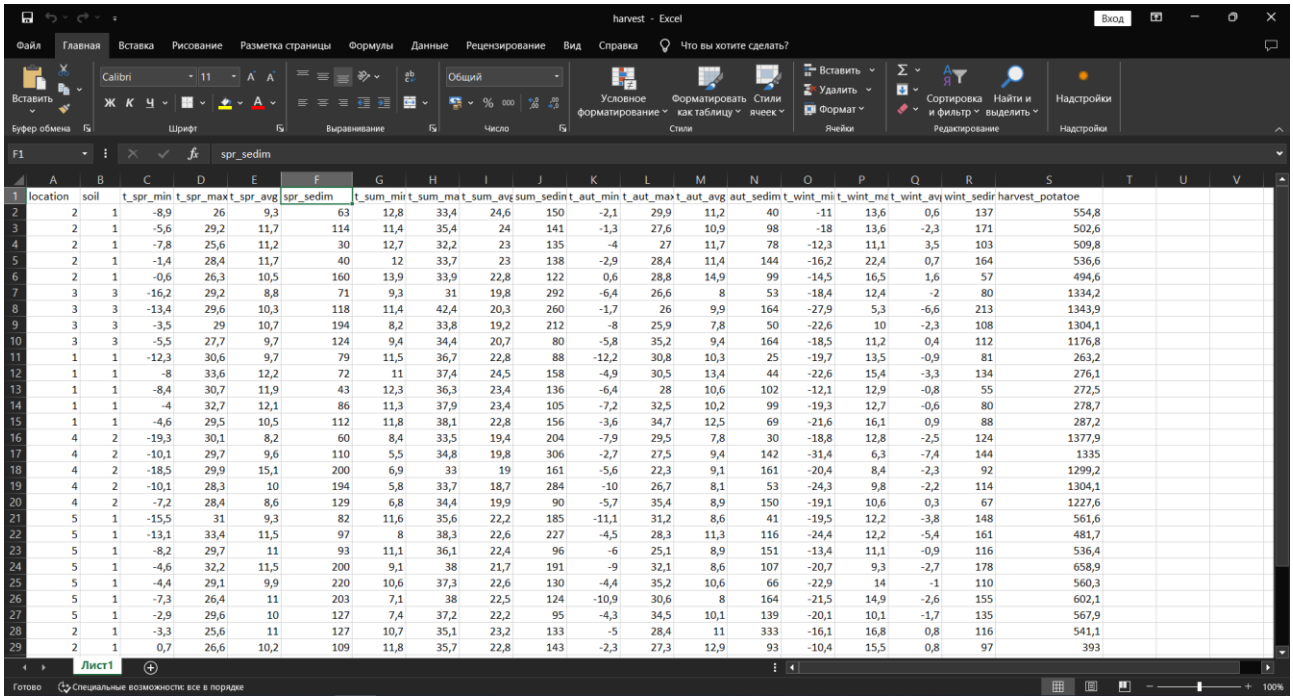
Залежно від конкретних обставин дослідження, можливо визначити числові значення ваг або використовувати категорії ваг для класифікації прикладів. Цей підхід дозволить враховувати контекст і значущість кожного прикладу у аналізі прогнозування врожаю.

4 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ ТА ПЕРЕВІРКА РЕЗУЛЬТАТІВ

4.1 Проведення дослідження

На підставі проведеного аналізу методів машинного навчання була вибрана модель випадкових лісів. З метою перевірки ефективності використання цієї моделі ми прийняли рішення створити датасет, використовуючи дані за період з 2011 по 2017 рік (рисунок 4.1), та конвертувати відповідні атрибути в числові параметри. Далі ми плануємо провести процес навчання на цьому датасеті, щоб отримати прогноз на наступні роки.

Цей підхід дозволяє нам не лише зрозуміти, наскільки актуально використання моделі випадкових лісів для прогнозування урожайності, але й отримати конкретні результати, які можуть слугувати основою для подальших висновків і вдосконалення методів прогнозування.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	location	soil	t_spr_min	t_spr_max	t_spr_avg	spr_sedim	t_sum_min	t_sum_max	t_sum_avg	sum_sedim	t_aut_min	t_aut_max	t_aut_avg	aut_sedim	t_wint_min	t_wint_max	t_wint_avg	wint_sedir	harvest_potatoe
2	2	1	-8,9	26	9,3	63	12,8	33,4	24,6	150	-2,1	29,9	11,2	40	-11	13,6	0,6	137	554,8
3	2	1	-5,6	29,2	11,7	114	11,4	35,4	24	141	-1,3	27,6	10,9	98	-18	13,6	-2,3	171	502,6
4	2	1	-7,8	25,6	11,2	30	12,7	32,2	23	135	-4	27	11,7	78	-12,3	11,1	3,5	103	509,8
5	2	1	-1,4	28,4	11,7	40	12	33,7	23	138	-2,9	28,4	11,4	144	-16,2	22,4	0,7	164	536,6
6	2	1	-0,6	26,3	10,5	160	13,9	33,9	22,8	122	0,6	28,8	14,9	99	-14,5	16,5	1,6	57	494,6
7	3	3	-16,2	29,2	8,8	71	9,3	31	19,8	292	-6,4	26,6	8	53	-18,4	12,4	-2	80	1334,2
8	3	3	-13,4	29,6	10,3	118	11,4	42,4	20,3	260	-1,7	26	9,9	164	-27,9	5,3	-6,6	213	1343,9
9	3	3	-3,5	29	10,7	194	8,2	33,8	19,2	212	-8	25,9	7,8	50	-22,6	10	-2,3	108	1304,1
10	3	3	-5,5	27,7	9,7	124	9,4	34,4	20,7	80	-5,8	35,2	9,4	164	-18,5	11,2	0,4	112	1176,8
11	1	1	-12,3	30,6	9,7	79	11,5	36,7	22,8	88	-12,2	30,8	10,3	25	-19,7	13,5	-0,9	81	263,2
12	1	1	-8	33,6	12,2	72	11	37,4	24,5	158	-4,9	30,5	13,4	44	-22,6	15,4	-3,3	134	276,1
13	1	1	-8,4	30,7	11,9	43	12,3	36,3	23,4	136	-6,4	28	10,6	102	-12,1	12,9	-0,8	55	272,5
14	1	1	-4	32,7	12,1	86	11,3	37,9	23,4	105	-7,2	32,5	10,2	99	-19,3	12,7	-0,6	80	278,7
15	1	1	-4,6	29,5	10,5	112	11,8	38,1	22,8	156	-3,6	34,7	12,5	69	-21,6	16,1	0,9	88	287,2
16	4	2	-19,3	30,1	8,2	60	8,4	33,5	19,4	204	-7,9	29,5	7,8	30	-18,8	12,8	-2,5	124	1377,9
17	4	2	-10,1	29,7	9,6	110	5,5	34,8	19,8	306	-2,7	27,5	9,4	142	-31,4	6,3	-7,4	144	1335
18	4	2	-18,5	29,9	15,1	200	6,9	33	19	161	-5,6	22,3	9,1	161	-20,4	8,4	-2,3	92	1299,2
19	4	2	-10,1	28,3	10	194	5,8	33,7	18,7	284	-10	26,7	8,1	53	-24,3	9,8	-2,2	114	1304,1
20	4	2	-7,2	28,4	8,6	129	6,8	34,4	19,9	90	-5,7	35,4	8,9	150	-19,1	10,6	0,3	67	1227,6
21	5	1	-15,5	31	9,3	82	11,6	35,6	22,2	185	-11,1	31,2	8,6	41	-19,5	12,2	-3,8	148	561,6
22	5	1	-13,1	33,4	11,5	97	8	38,3	22,6	227	-4,5	28,3	11,3	116	-24,4	12,2	-5,4	161	481,7
23	5	1	-8,2	29,7	11	93	11,1	36,1	22,4	96	-6	25,1	8,9	151	-13,4	11,1	-0,9	116	536,4
24	5	1	-4,6	32,2	11,5	200	9,1	38	21,7	191	-9	32,1	8,6	107	-20,7	9,3	-2,7	178	658,9
25	5	1	-4,4	29,1	9,9	220	10,6	37,3	22,6	130	-4,4	35,2	10,6	66	-22,9	14	-1	110	560,3
26	5	1	-7,3	26,4	11	203	7,1	38	22,5	124	-10,9	30,6	8	164	-21,5	14,9	-2,6	155	602,1
27	5	1	-2,9	29,6	10	127	7,4	37,2	22,2	95	-4,3	34,5	10,1	139	-20,1	10,1	-1,7	135	567,9
28	2	1	-3,3	25,6	11	127	10,7	35,1	23,2	133	-5	28,4	11	333	-16,1	16,8	0,8	116	541,1
29	2	1	0,7	26,6	10,2	109	11,8	35,7	22,8	143	-2,3	27,3	12,9	93	-10,4	15,5	0,8	97	393

Рисунок – 4.1 Невідформатований датасет

Використання реальних даних та їх подальший аналіз стануть ключовим етапом у дослідженні та перевірці ефективності обраної моделі. Тому ми використовуємо існуючі метеорологічні та геологічні данні періоду з 2018 по 2021 рік для ознак та на їх основі складаємо прогноз (рисунок 4.2).

```
D:\KhNURE\M\M-t1\PyProj\RandomForest\.venv\Scripts\python.exe D:\KhNURE\M\M-t1\PyProj\RandomForest\main.py
   location  soil  t_spr_min  ...  t_wint_avg  wint_sedim  harvest_potatoe
0         2     1       -8.9  ...         0.6       137.0            554.8
1         2     1       -5.6  ...        -2.3       171.0            502.6
2         2     1       -7.8  ...         3.5       103.0            509.8
3         2     1       -1.4  ...         0.7       164.0            536.6
4         2     1       -0.6  ...         1.6        57.0            494.6

[5 rows x 19 columns]
Введіть дані для прогнозу:
Місцезнаходження (Kherson(1)/Odesa(2)/Zhytomyr(3)/Rivne(4)/Dnipro(5)): 1
Тип грунту (black_soil(1)/sod_podzolic(2)/gray_forest(3)): 1
Мінімальна температура весною: -14.3
Максимальна температура весною:31
Середня температура весною:11.7
Опади за весну:100
Мінімальна температура літом: 10.2
Максимальна температура літом: 36.9
Середня температура літом: 24.2
Опади за літом: 114
Мінімальна температура восени: -7.3
Максимальна температура восени: 35
Середня температура восени: 11.6
Опади за осінь: 84
Мінімальна температура взимку: -14.5
Максимальна температура взимку: 11.8
Середня температура взимку: -0.2
Опади за зиму: 114
Прогноз урожайності: 282.98900000000003
```

Рисунок – 4.2 Робота программного застосунку

4.2 Аналіз результатів, оцінка, рекомендації

Прогноз урожайності картоплі був розроблений, використовуючи відомості про процеси вирощування цієї культури в ґрунтах типу чорнозему в Херсонській області. Отримані результати прогнозу були систематизовані та введені до таблиці 4.1.

Таблиця 4.1 – Порівняння реального врожаю з прогнозованим

Рік	Врожай (тис. т.)		Тенденція	Похибка	
	Прогнозований (тис .т)	Реальний (тис .т)		Абсолютна	Відносна
2018	282.9	295.2	Співпадає	-12,1	4,1%
2019	287,3	254,7	Не співпадає	32,6	12,8%
2020	288,7	281,2	Співпадає	7,5	2,7%
2021	355,1	426,5	Співпадає	71,4	16,7%

Для більш зручного сприйняття було побудовано графік, що відображено на рисунку 4.3.

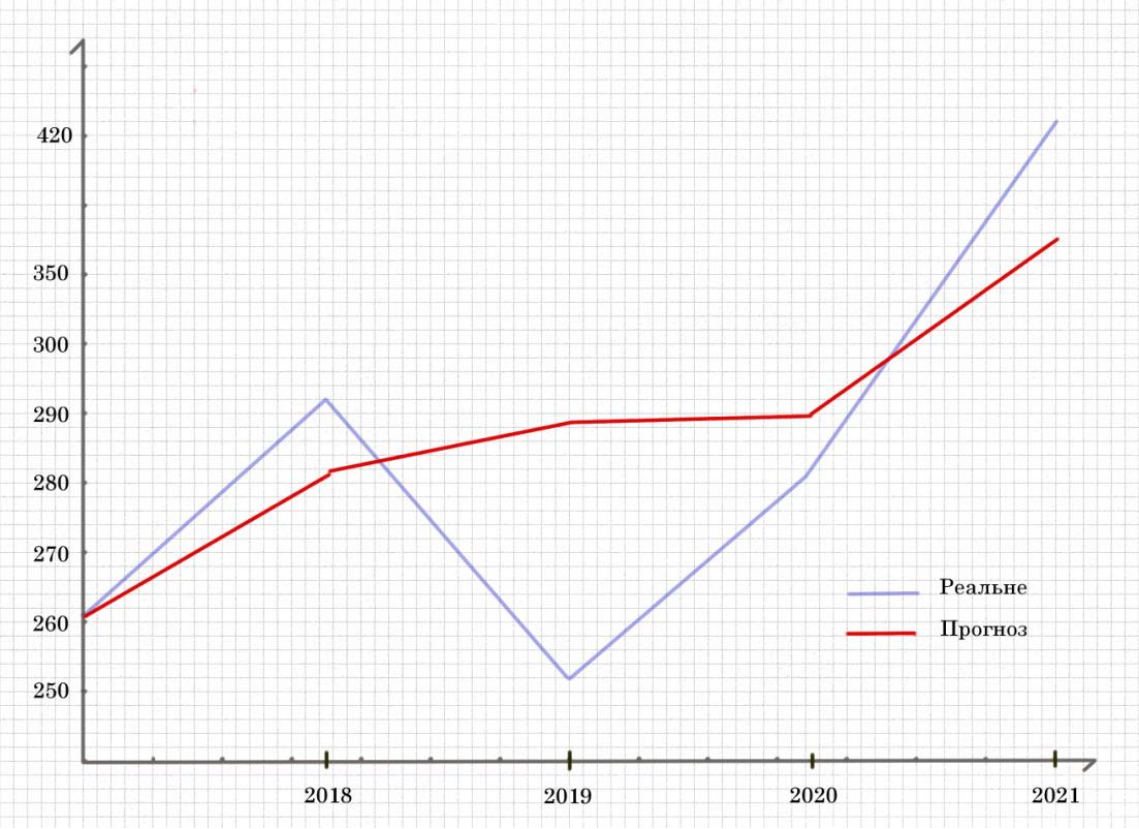


Рисунок – 4.3 Порівняльний графік врожайності

Як можна спостерігати, основні тенденції росту та спаду врожайності прогноз відображає вірно. Проте, важливо зауважити, що найбільші похибки

спостерігаються у тих роках, коли врожайність значно відрізняється в порівнянні з іншими періодами.

Є кілька можливих причин таких великих коливань у врожайності, які можуть впливати на точність прогнозів. Однією з можливих причин є непередбачені природні явища, такі як погодні катастрофи або екстремальні кліматичні умови, які можуть суттєво вплинути на вирощування сільськогосподарських культур. Окрім цього, варто зауважити, що в датасеті фіксувались коливання що сезонно, що може бути критичним при їх виявлені, що сильно впливає на прогноз врожайності. Щоб знизити ризик погрішності в прогнозах урожайності, можна розглянути наступні підходи:

1. збільшення обсягу тренувальних даних: збільшення кількості даних для тренування моделі може поліпшити її здатність враховувати великі коливання, важливо включити різні роки з різними умовами, щоб модель могла засвоїти широкий спектр сценаріїв.

2. збільшення кількості характеристик: додаткові фактори, такі як кліматичні умови, ґрунтовий склад чи кислотність води, можуть внести важливий внесок у формування врожайності.

3. уточнення особливостей років з великими коливаннями: дослідження та ідентифікація конкретних факторів, які призводять до великих коливань у врожайності в певні роки, допоможе моделі краще адаптуватися до таких умов, наприклад, це може бути пов'язано зі специфічними погодними або географічними умовами.

4. оптимізація гіперпараметрів моделі [2]: проведення подальшого тюнінгу гіперпараметрів моделі може покращити її властивості прогнозування, дослідження важливості окремих параметрів та їх вплив на результати може бути корисним.

Ці заходи можуть сприяти зменшенню похибок у прогнозах, забезпечуючи більш надійні та точні результати, навіть у роках із значними коливаннями у вирощуванні сільськогосподарських культур.

4.3 Створення архітектури інформаційної системи

У даному підрозділі ми визначимо структуру та компоненти нашої майбутньої інформаційної системи, котра призначена для прогнозування врожайності за допомогою методів машинного навчання, зокрема, методу випадкових лісів.

Почнемо з чіткого визначення вимог до нашої системи. Ми проаналізуємо ключові функціональні та технічні характеристики, які гарантують ефективну та надійну роботу системи прогнозування врожайності та визначимо ряд основних бізнес-функцій. Цей етап є критичним для розробки системи, яка відповідає вимогам та очікуванням користувачів.

Другий аспект розділу розглядатиме вимоги до інтерфейсу користувача. Ми окреслимо необхідність та очікування від інтерфейсу, щоб забезпечити зручність користування та зрозумілість для кінцевого користувача. Також буде розроблено прототип інтерфейсу, щоб візуалізувати функціонал системи та забезпечити можливість взаємодії з користувачем.

Цей етап роботи визначає основні особливості та можливості нашого прогностичного інструменту, роблячи акцент на високу точність та зручність використання.

4.3.1 Визначення вимог та функцій інтерфейсу

Розробка функціональних вимог до інформаційної системи передбачає визначення основних бізнес-функцій, які відповідають потребам користувачів в контексті системи прогнозування врожайності для програмного застосунку.

Усього було виділено три категорії користувачів: «незареєстрований користувач», «користувач», «адміністратор». Під «незареєстрований користувач» ми розглядаємо особу що не увійшла в систему. Звичайний користувач, що пройшов процес реєстрації, за для використовування програмного продукту буде мати статус «користувач». «Адміністратором» ми

розглядаємо осіб, що будуть відповідати за технічну сторону продукту: вдосконалення, оновлення, тренування алгоритму, внесення нових даних, та за необхідності надання технічної підтримки «користувачу». Визначаємо функції для кожного з них:

Бізнес-функції системи для незареєстрованих користувачів:

1. вхід в систему для користувачів зі статусом "гість";
2. перегляд загальної статистичної інформації без можливості редагування, збереження даних та складення прогнозів;
3. обмежене виконання пошуку за параметрами року врожайності та типу культури;
4. реєстрація для переходу на статус «користувач».

Бізнес-функції системи для зареєстрованих користувачів:

1. вхід в систему з визначенням статусу «користувач»;
2. усі функції, що були доступні «незареєстрованому користувачу»;
3. повний доступ до виконання пошуку за параметрами року врожайності та типу культури, за доступні роки;
4. введення, редагування та збереження власних даних про поля, культури, сорти, типи ґрунтів, для подальшого прогнозу або обліку;
5. проведення прогнозу врожайності за введеними параметрами;
6. проведення аналізу впливу зовнішніх факторів на урожайність відповідно проведених прогнозів.

Бізнес-функції системи для адміністраторів:

1. вхід в систему з визначенням статусу «адміністратор»;
2. усі функції, що притаманні «користувачу»;
3. можливість додавання нових користувачів і надання їм доступу до системи;
4. контроль доступу та прав користувачів;
5. доповнення тренувальних датасетів;
6. моніторинг системи та підтримка її безперебійної роботи.

Ці бізнес-функції дозволяють визначити основні можливості інформаційної системи для прогнозування урожайності та структурувати їх для подальшої реалізації у вигляді програмного застосунку. Користувачі різних категорій матимуть доступ до різних функцій, що забезпечить ефективне та зручне використання системи в контексті прогнозування урожайності.

На основі отриманої інформації, ми визначаємо функції інтерфейсу клієнтської частини інформаційної системи. Створюємо компактний опис, узявши до уваги, що функціональна частина реалізується на стороні сервера баз даних (MySQL). Результати більш детально відображені у таблиці 5.4.

Таблиця 4.2 – Перелік елементів інтерфейсу й бізнес-функцій

№	Елементи інтерфейсу	Бізнес-функція інформаційної системи
1	Вікно входу у систему: вікно реєстрації / авторизації; кнопки «реєстрація», «вхід»; текстове поле статусу.	Виконується для реєстрації або авторизації. Кнопки дозволяють зареєструватись щоб (або) увійти в систему відповідно, та отримати статус «користувач». Інакше, користувач залишається у статусі «незареєстрований користувач» і має обмежений доступ до додатку.
2	Вікно користувача.	Відображає данні користувача. Його попередньо складені прогнози та інформацію о ділянках для швидкого імпорту до прогнозів.
3	Головна сторінка, що містить кнопки переходу до архіву, вводу власної ділянки та самого прогнозу.	Головна сторінка виконує навігацію до основних бізнес-функцій. Серед котрих лише до часткового перегляду архіву надано доступ «незареєстрованому користувачу». Для подальшого необхідно увійти в систему.

Продовження таблиці 4.2

№	Елементи інтерфейсу	Бізнес-функція інформаційної системи
4	Архів: пошукове вікно вводу інформації, кнопка «детальніше», вікно інформації.	Використовується для власного перегляду інформації минулих років, та базової інформації щодо циклу зростання прогнозованих рослин.
5	Вікно заповнення інформації ділянки, кнопка «закріпити».	Вікно для вводу власної ділянки, як об’єкту прогнозу.
6.	Вікно заповнення для прогнозу, кнопка «спрогнозувати».	Вікно для вводу основних даних для прогнозу.
7	Вікно прогнозу, кнопка «зберегти»	Виведений прогноз з відповідною інфографікою, зберігається у профілі користувача.

При проектуванні інтерфейсу користувача, кожен крок є ключовим для створення зручного та ефективного взаємодії з інформаційною системою. User Stories (історії користувача) [28] виступають важливим інструментом, що допомагає визначити функціональність та очікування кінцевого користувача. Ці історії спрямовані на створення продукту, який не лише відповідає технічним вимогам, але й враховує потреби та зручність використання для кінцевого користувача.

User Story – це простий і швидкий спосіб документування вимог клієнта без необхідності створення об’ємних формалізованих документів. Вони створюються в рамках завершених послідовних подій за концепцією: початкові умови, коли, тоді, коли, тоді. Це дозволяє чітко визначити, що робить користувач та як система повинна реагувати.

Структурований опис проекту дозволяє замовнику зосередитися на важливих елементах та контролювати відповідність кінцевого продукту вимогам. Він також допомагає розробникам оцінити об’єм, розділити функціонал на версії продукту та концентруватися на конкретних задачах. Зміни вимог можуть бути легко внесені, що полегшує управління та аналіз проекту.

User Stories наводяться для детального опису кожного варіанту використання з контекстної Use Case-діаграми у таблицях нижче.

Таблиця 4.3 – UserStory «Вхід у систему».

Дійові особи	Адміністратор, Незареєстрований користувач, Система.
1	2
Цілі	Незареєстрований користувач: увійти в систему. Система: зміна статусу користувача.
Успішний сценарій: 1. Незареєстрований користувач вводить дані авторизації. 2. Система змінює статус на «Користувач». Користувач входить у власний профіль.	
Результат	Авторизація користувача виконана.
Розширення:	
1	2
*a	Незареєстрований користувач не проходив реєстрацію. Результат: Система повідомляє, що дані користувача відсутні.

Продовження таблиці 4.3.

1	2
*б	Незареєстрованому користувачу повідомляють, що його даних не має у системі. Користувач звертається до адміністрації. Результат: Адміністрація отримує запит на відновлення акаунту.

Таблиця 4.4 – UserStory «Реєстр ділянки».

Дійові особи	Адміністратор, Користувач, Система
1	2
Цілі	Користувач: ввести інформацію про ділянку. Система: перевірити на та ввести інформацію у датасети. Адміністратор: підтвердити коректність роботи застосунку
Успішний сценарій: 1. Користувач переходить до форми внесення даних ділянки. 2. Система відображає йому форму для заповнення гео та метеоданих. 3. Користувач вносить свої данні, зберігає заповнену форму. 4. Система перевіряє на коректність заповнені поля форми. 5. Система вносить дані у датасети. 6. Користувач закріплює дані ділянки у своїх шаблонах. 7. Система повідомляє користувачу, що реєстрація ділянки успішно оформлена та шаблон закріплено.	
Результат	Успішно оформлено замовлення на користувача.

Продовження таблиці 4.4.

1	2
Розширення:	
*а	Система повідомляє що дані локації введені некоректно. Очищується поле з геолокацією.
1а	Користувач визначив заповнення полів помилковим. Результат: Користувач знову заповнює форму.
2а	Користувач повідомив Адміністрації, що в заповненні полів помилки. Результат: Адміністрація вносить правки в можливість заповнення форми.

Таблиця 4.5 – UserStory роботи «Обмежений перегляд архіву»

Дійові особи	Адміністратор, Незареєстрований користувач, Система.
1	2
Цілі	Незареєстрований користувач: переглянути інформацію в архіві. Система: обмежити доступ до інформації.
Успішний сценарій: 1. Незареєстрований користувач відкриває пошукову панель архіву та вводить запит. 2. Система дає доступ до інформації архіву. 3. Користувач отримує бажану інформацію.	

Продовження таблиці 4.5

Результат	Запит на інформацію підтверджено, незареєстрований користувач отримує інформацію.
Розширення:	
1	2
*а	Незареєстрований користувач запитує дані перевищуючі його доступ. Результат: Система повідомляє, що недостатньо користувацьких вповноважень.
*б	Незареєстрованому користувачу повідомляють, що за запитом інформації немає. Результат: Незареєстрований користувач знову заповнює поля пошуку.
1а	Незареєстрований користувач проводить вхід у систему. Результат: Незареєстрований користувач змінює статус і отримує доступ до інформації
1б	Незареєстрований користувач робить запит до Адміністрації на оновлення даних. Результат: Адміністрація отримує запит на оновлення даних.

Таблиця 4.6 – UserStory роботи «Прогнозування врожайності»

Дійові особи	Користувач, Система
1	2
Цілі	Користувач: отримати прогноз врожайності. Система: провести прогноз з використанням машинного навчання, вивести та зберегти результат користувачу.
Успішний сценарій: 1. Користувач заповнює поля інформації для прогнозу. 2. Система генерує прогноз. 3. Система виводить користувачу спрогнозовану інформацію. 4. Система додає прогноз в історію прогнозів користувача.	
Результат	Прогноз врожайності отримано і збережено.
1	2
Розширення:	
*а	Користувач ввів некоректні дані для прогнозу. Результат: Видало помилку, користувач знову заповнює форми.

4.3.2 Візуальна складова інтерфейсу

Для ефективної розробки інтерфейсу користувача для нашого додатку, який забезпечує прогнозування урожайності, перш за все, потрібно чітко визначити напрямок роботи та налаштувати його так, щоб задовольнити потреби

максимальної кількості користувачів. Для досягнення цього ми проведемо аналіз та визначимо цільову аудиторію.

Цільова аудиторія нашого додатку - це фермери, агрономи та спеціалісти в галузі сільського господарства, які зацікавлені у точних та передбачуваних прогнозах урожайності, окрім його можуть бути зацікавленими інвестори та люди професій що контактують з аграрним сектором. Наш інструмент дозволить їм отримувати точні та актуальні дані про прогноз врожайності, а також здійснювати додаткові аналітичні операції.

У сучасних умовах наша система допоможе максимально використовувати доступні дані та знижувати ризики для аграріїв. Користувачі зможуть здійснювати прогнозування в один дотик та отримувати необхідну інформацію щодо урожайності.

Вік середнього користувача 30-35 років. Стать не має значення, однак серед користувачів переважно чоловіки. Основна мова користування - українська. Рід занять потенційних клієнтів переважно агрономи, підприємці в аграрній сфері, садоводи рідше можуть бути працівники інших сфер що просто зацікавленні кімнатними рослинами (так як вони теж є частиною асортименту).

Враховуючи інформацію про портрети користувачів програмного додатку, було визначено стильові рішення, і обрано варіант, спрямований на досягнення мінімалізму. Мінімалізм широко використовується в різних галузях людської діяльності і визначається як стиль або техніка, відзначена обмеженою кількістю елементів та простотою. Зберігаючи базові характеристики - простоту та смислову насиченість, мінімалізм вирізняється ключовим принципом: залишати лише найнеобхідніші елементи для створення враження елегантності. Лінії, фігури, точки, кольори, порожній простір і композиція служать визначеній меті, дотримуючись ретельної організації. Мінімалізм в сучасному світі застосовується в різних сферах, таких як архітектура, живопис, фотографія, дизайн і навіть сервіровки.

Мінімалізм у дизайні виражається через:

- простоту та ясність;
- виражену візуальну ієрархію;
- увагу до пропорцій і композиції;
- функціональність кожного елемента;
- використання порожнього простору;
- відсутність зайвих деталей, з фокусом на кожному елементі.

Мінімалізм робить інтерфейси зручнішими для користувача, наголошуючи на важливих елементах і роблячи шлях користувача інтуїтивним. Мінімалістичні інтерфейси виглядають вишукано і акуратно, що додає естетичне задоволення та привабливість у UX.

Було вирішено розробляти мінімалістичний дизайн з обмеженою палітрою кольорів, що відображена на рисунку 4.4 і в той же час з зрозумілим і зручним у використанні інтерфейсом.

Рисунок 4.4 – Підбор кольорової палітри

Далі було створено ескіз інтерфейсу у кольорі та з дотриманням обраного стилю:

- сторінку входу у систему (рис. 4.5);
- сторінку профілю (рис. 4.6);
- головну сторінку (рис 4.7);
- сторінка виведення прогнозів (рис. 4.8).

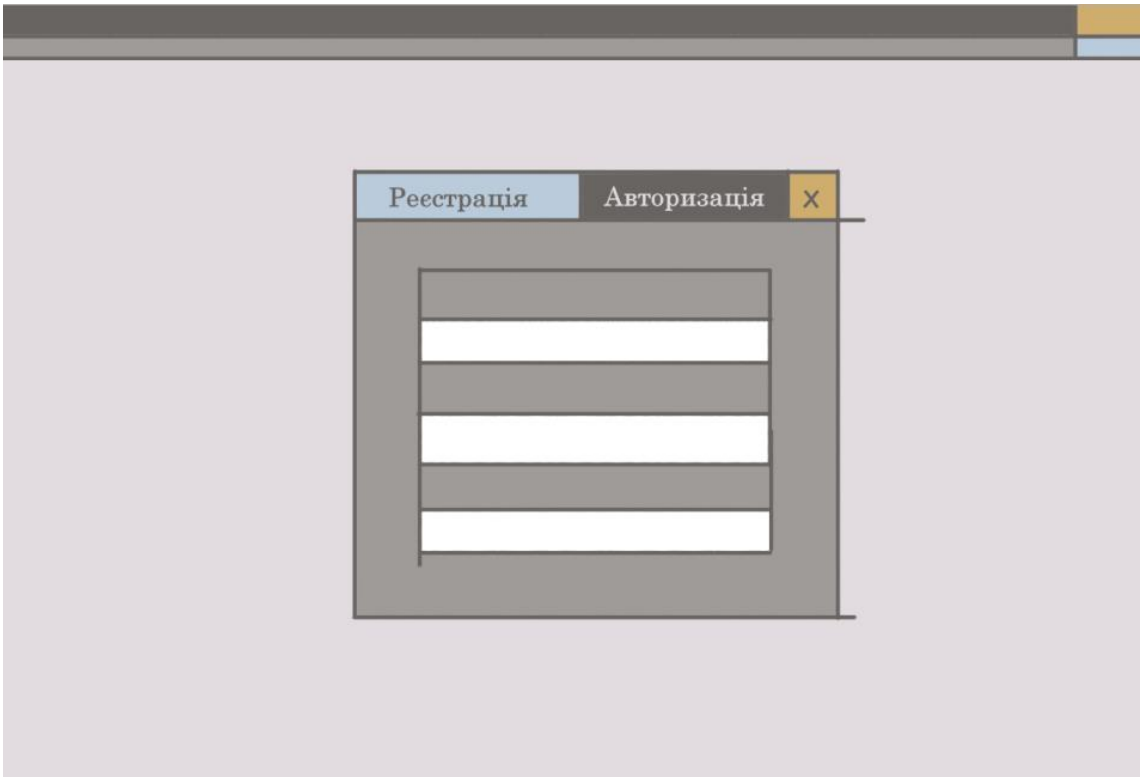


Рисунок 4.5 – Прототип сторінки входу у систему

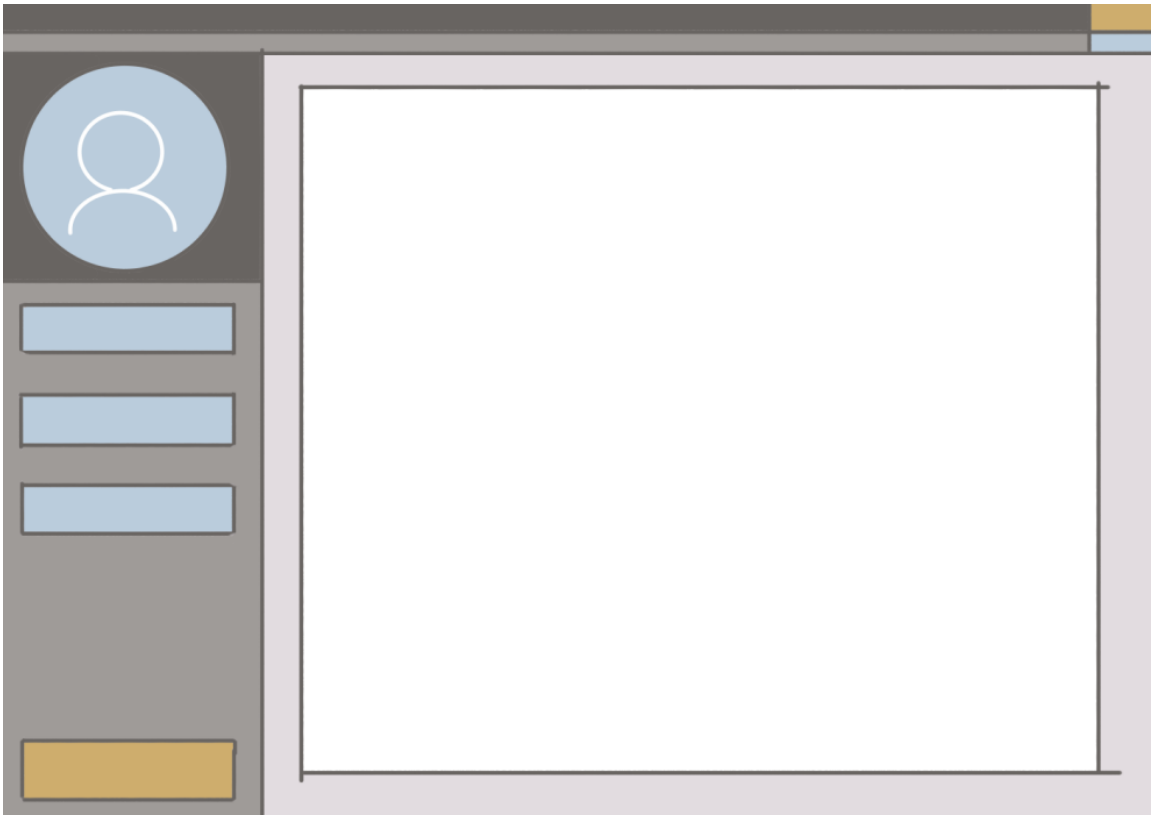


Рисунок 4.6 – Прототип профілю

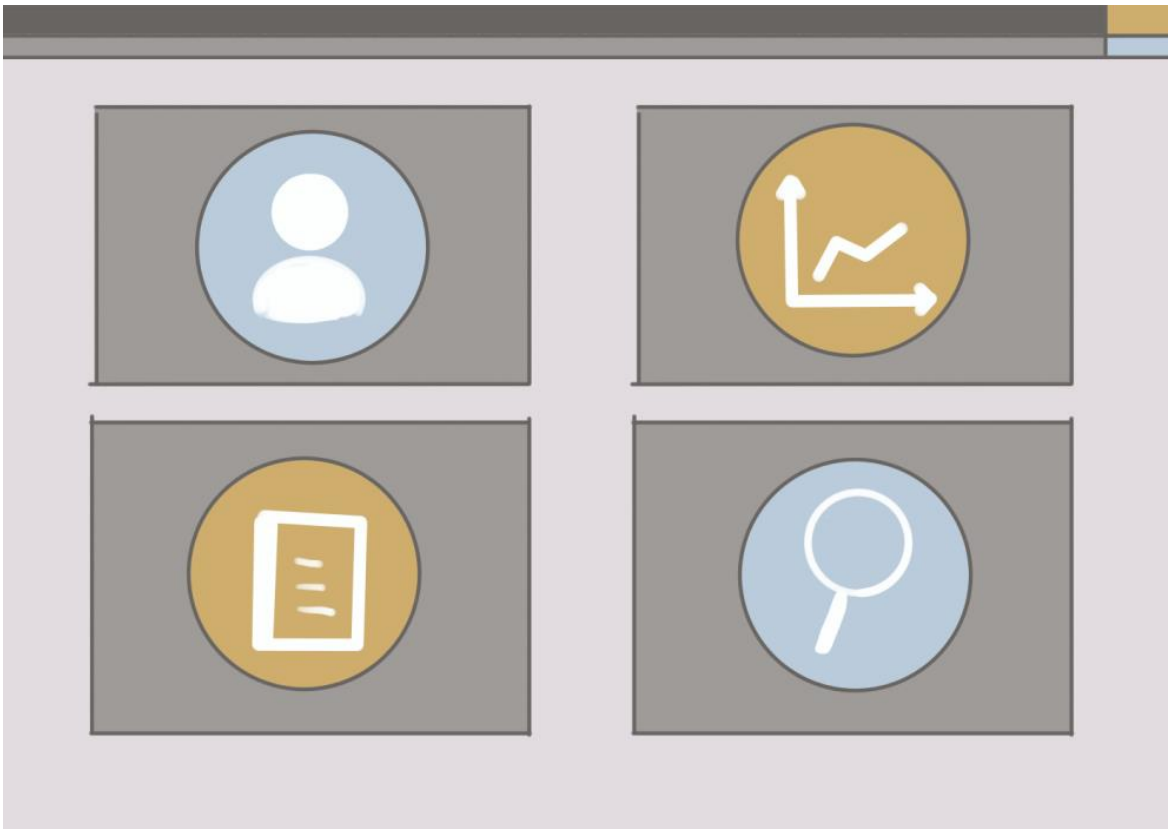


Рисунок 4.7 – Прототип головної сторінки

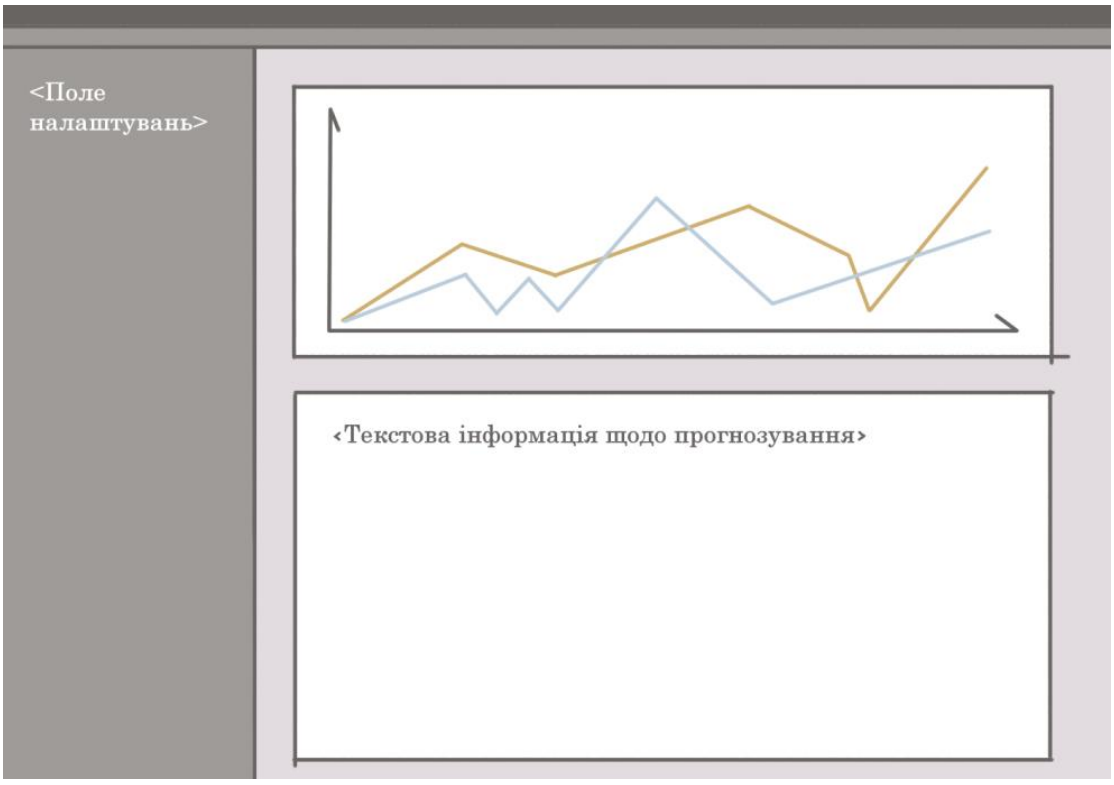


Рисунок 4.8 – Прототип сторінки прогнозів

І не менш важливо розуміти що усі сторінки повинні мати той чи інший зв'язок між собою, забезпечувати перехід від однієї сторінки к іншій маючи при цьому певну логіку. Саме за для цього була побудована карта навігації по додатку, що зображена на рисунку 4.9.

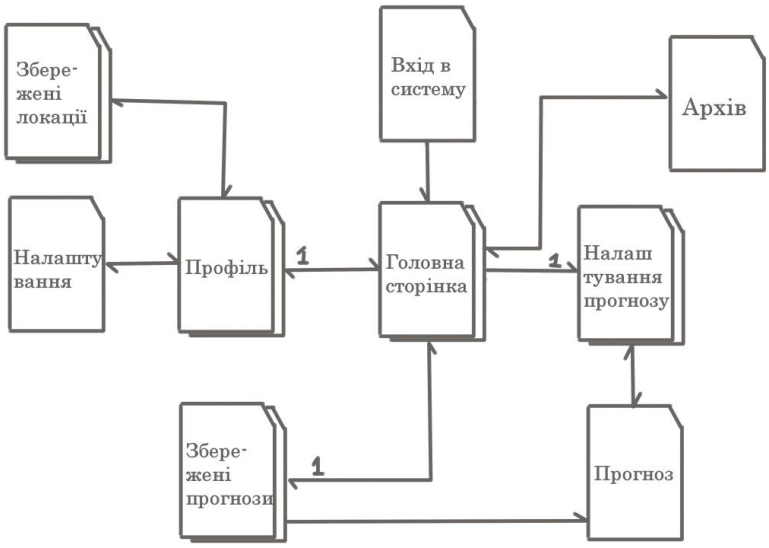


Рисунок 4.9 – Карта навігації по додатку.

ВИСНОВКИ

В ході виконання даного дослідження було проведено аналіз предметної області, спрямований на вивчення можливостей прогнозування врожайності сільськогосподарських культур з використанням методів машинного навчання. Сформульована задача, що ставить за мету поліпшення прогнозування врожайності та оптимізації агропромислового сектору.

Для вирішення поставленої задачі вибрано різні методи машинного навчання, зокрема методи часових рядів, лінійну регресію та метод опорних векторів. Здійснено порівняльний аналіз цих методів, визначено їх переваги та недоліки, а також розглянуто можливості вдосконалення застосування методу випадкових лісів, включаючи додавання ваг до прикладів.

Проведено дослідження роботи методу випадкових лісів для прогнозування врожайності сільськогосподарських культур. Виявили складні для прогнозування випадки та визначили аспекти для покращення роботи методу.

Також, розроблена архітектура системи, що враховує особливості обраного методу та забезпечує ефективне впровадження моделі прогнозування. Додатково, розроблено прототип інтерфейсу, який відображає зручний та інтуїтивно зрозумілий спосіб взаємодії з системою.

Висновок з даного дослідження полягає в тому, що використання методів машинного навчання є багатообіцяючим напрямком для покращення прогнозування врожайності у сільському господарстві. Вибір конкретного методу повинен бути обґрунтованим з урахуванням особливостей задачі та доступних даних. Розгляд вагомості кожного методу та його можливостей у контексті покращення сільськогосподарської продуктивності є ключовим аспектом подальших досліджень та впровадження в практику.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Дослідження Методів Машинного Навчання для Прогнозування / Овчаренко А., Петрова Р. // Інформаційні системи та технології: матеріали 12-ї Міжнародної науково-технічної конференції. Частина 2. Молодіжна секція, Харків, 28 листопада 2028 – 01 грудня 2023 року / наук. ред. В.В. Безкоровайний, L. Petryshyn, З.В. Дудар, Ю.В. Міщеряков. – Х.: ХНУРЕ, 2023. – С. 51-52

2. J. Nilsson N. Introduction to machine learning : підручник. Stanford, CA 94305, 2005. 179 p.

3. Комплекс навчально-методичного забезпечення навчальної дисципліни «Комп`ютерний зір» підготовки магістра спеціальності 122 - Комп’ютерні науки / ХНУРЕ ; розроб. О. Г. Аврунін. – Харків, 2022. – 381 с.

4. Рослинництво: Підручник / О. І. Зінченко, В. Н. Салатенко, М. А. Білоножко; За ред. О. І. Зінченка. — К.: Аграрна освіта, 2001. — 591 с.

5. Прогнозування і програмування врожаїв сільськогосподарських культур: для виконання практичних завдань студентами факультету агрономії за спеціальністю 201 Агрономія. - Умань, 2019 р.— Умань: Редакційновидавничий відділ УНУС, 2018. — 40 с.

6. CISS GROUP, Прогнозування майбутнього врожаю. https:// ciss-group.com. URL: https://ciss-group.com/ua/poslugi/syurvej-ukr/syurvej-po-galuzyam/agrarna-promislovist/inshi-inspekcziї-silskogospodarskoї-produkcziї-ta-poslugi/436-prognozuvannya-majbutnogo-vrozhayu.html (дата звернення: 30.12.2023)

7. EOS Data Analytics. https://eos.com. URL: https://eos.com/uk/ products/crop-monitoring/custom-solutions/yield-prediction/ (дата звернення: 30.12.2023)

8. METOS, https://metos.at. URL: https://metos.at /ua/yield-prediction/ (дата звернення: 30.12.2023)

9. Машинне навчання простими словами. Частина 1. http://www.mmf.lnu.edu.ua. URL: http://www.mmf.lnu.edu.ua/ar/1739 (дата звернення: 27.12.2023).

10. Забродоцька Л.Ю. Основи агрономії : навчальний посібник / Л.Ю. Забродоцька. – Луцьк : Інформ.-вид. відділ Луцького НТУ, 2019. – 360 с

11. Державна служба статистики України. https://www.ukrstat.gov.ua. URL: https://www.ukrstat.gov.ua/ (дата звернення: 29.10.2023).

12. Хомик Н. І. Основи агрономії: навчальний посібник до практичних занять та самостійної роботи / Н. І. Хомик, Г. Б. Цьонь, Т. А. Довбуш, Н. А. Антончак. – Тернопіль: ФОП Паляниця В. А., 2021. – 320 с.

13. SuperArronom, карта гуртів України. https://superagronom.com URL: https://superagronom.com/karty/karta-gruntiv-ukrainy (дата звернення: 28.10.2023).

14. Архів погоди Kyiv - meteoblue. https://www.meteoblue.com URL: https://www.meteoblue.com/uk/weather/historyclimate/weatherarchive/kyiv_ukraine_703448?fcstlength=1m&year=2022&month=11 (дата звернення: 28.10.2023).

15. Комплекс навчально-методичного забезпечення навчальної дисципліни "Машинне навчання", підготовки бакалавра, напрям 6.050101 - Комп’ютерні науки [Електронний ресурс] / ХНУРЕ ; розроб. О. В. Вітько. – Харків, 2017. – 265 с.

16. Chollet F. Deep Learning with Python. Shelter island : Manning Publications Co., 2018. 361 p.

17. Shmueli G., Lichtendahl Jr K. C., Practical time series forecasting with R a hands-on guide, 2nd ed. axelrod schnall publishers, 2016

18. ARIMA modelling in R, https://otexts.com URL: https://otexts.com/fpp2/arima-r.html (дата звернення: 05.01.24)

19. Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(1), 1–22.

20. Burges C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery 2:121–167, 1998

21. Hastie T., Tibshirani R., Friedman J., The elements of statistical learning data mining, inference, and prediction, 12th ed., Springer, 2017.

22. Ben-Hur A. Support vector clustering. Journal of machine learning research. 2001. Vol. 2. P. 125–137.

23. КНУ імені Тараса Шевченка, Метод опорних векторів, 7 лекція http://om.univ.kiev.ua URL: http://om.univ.kiev.ua/users_upload /15/upload/file/pr_lecture_07.pdf (дата звернення: 02.01.24)

24. Бречко Д. О., Максишко Н. К., Іванов С. М. Інтелектуальний аналіз даних : конспект лекцій. Запоріжжя : ЗНУ, 2020. 156 с.

25. Random forests - classification description. Statistics at UC Berkeley | Department of Statistics. URL: https://www.stat.berkeley.edu/~breiman /RandomForests/cc_home.htm (date of access: 14.11.2023).

26. Breiman L. Machine Learning. 2001. Vol. 45, no. 1. P. 5–32. URL: https://doi.org/10.1023/a:1010933404324 (date of access: 09.01.2024).

27. https://hashdork.com URL: https://hashdork.com/uk/алгоритми-машинного-навчання-для-новачків/ (06.01.2024).

28. QATestLab Training center, що таке user story і як її писати https://training.qatestlab.com URL: https://training.qatestlab.com/blog/technical-articles/user-story/ (дата звернення: 20.12.24)

29. Information Technology Based on Qualitative Methods in Cyber-Physical Systems of Situational Disaster Risk Management / Grebennik I., Hutsa O., Petrova R., Yelchaninov D., Morozova A. // In: Murayama Y., Velev D., Zlateva P. (eds) Information Technology in Disaster Risk Reduction. ITDRR 2020. IFIP Advances in Information and Communication Technology, vol 622. Springer, Cham. https://doi.org/10.1007/978-3-030-81469-4_11, pp 132-143

30. ДСТУ 3008:2015. Інформація та документація. Звіти у сфері науки і техніки. Структура та правила оформлювання / Нац. стандарт України. – Вид. офіц. – [Чинний від 2017-07-01]. – Київ: ДП «УкрНДНЦ», 2016. – 26 с.

31. ДСТУ 7.1:2006. Система стандартів з інформації, бібліотечної та видавничої справи. Бібліографічний запис. Бібліографічний опис. Загальні вимоги та правила складання / Нац. стандарт України. – Вид. офіц. – [Чинний від 2007-07-01]. – Київ : Держспоживстандарт України, 2007. – 47 с.

32. Методичні вказівки до організації виконання та захисту кваліфікаційної роботи на здобуття другого (магістерського) рівня вищої освіти спеціальності 122 Комп'ютерні науки, освітньо-професійна програма «Інформаційні технології проєктування» / Упорядники: І.В. Гребеннік, — В.Г. Іванов, А.1. Коваленко, О.Б. Колесник, Ю.В. Міщеряков, І.А. Урняєва, С.І. Чайніков. Харків: ХНУРЕ, 2021. 54 с.