НАЦІОНАЛЬНИЙ УНІВЕРСТИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Факультет інформаційних технологій

ПОГОДЖЕНО
Декан факультету

Інформаційних технологій
(назва факультету)

_____ Ігор Болбот
(підпис) (ім'я прізвище)

"___"_____2025 р.

ДОПУСКАЄТЬСЯ ДО ЗАХИСТУ
Завідувач кафедри

комп'ютерних наук
(назва кафедри)

_____ Белла Голуб
(підпис) (ім'я прізвище)

"___"_____2025 р.

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему Система інтелектуального аналізу параметрів вирощування агрокультур

Спеціальність 122 «Комп'ютерні науки»
(Код і найменування)

Освітня програма Інформаційні управляючі системи та технології
(Назва)

Орієнтація освітньої програми Освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Гарант освітньої програми

к.т.н., доцент _____ Белла Голуб
(науковий ступінь та вчене звання) (підпис) (ім'я прізвище)

Керівник магістерської кваліфікаційної роботи

к.т.н., доцент _____ Белла Голуб
(науковий ступінь та вчене звання) (підпис) (ім'я прізвище)

Виконав _____ Олександр Пухальський
(підпис) (ім'я прізвище студента)

НАЦІОНАЛЬНИЙ УНІВЕРСТИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Факультет (ННІ) інформаційних технологій

ЗАТВЕРДЖУЮ
Завідувач кафедри комп'ютерних наук
доцент, к.т.н. _____ Белла Голуб
(науковий ступінь, вчене звання) (підпис) (ім'я прізвище)
"01" листопада 2024 року

З А В Д А Н Н Я

ДО ВИКОНАННЯ МАГІСТЕРСЬКОЇ КВАЛІФІКАЦІЙНОЇ РОБОТИ СТУДЕНТУ

Пухальському Олександру Вадимовичу
(прізвище, ім'я, по батькові)

Спеціальність 122 «Комп'ютерні науки»
(код і назва)

Освітня програма Інформаційні управляючі системи та технології
(назва)

Орієнтація освітньої програми освітньо-професійна

Тема магістерської кваліфікаційної роботи Система інтелектуального аналізу параметрів вирощування агрокультур
затверджена наказом ректора НУБіП України від "01" листопада 2024р. №1964 «С»
Термін подання завершеної роботи на кафедру 20.11.2025
(рік, місяць, число)
Вихідні дані до магістерської кваліфікаційної роботи: набори даних відкритого доступу (державна агростатистика, кліматичні та грунтов дані)
Перелік питань, що підлягають дослідженню:

1.Виокремлення та опис ключових агроекологічних змінних, що зумовлюють коливання врожайності.

2. Проєктування та використання OLAP і методів Data Mining для підтримки рішень у рослинництві та підвищення ефективності управління ресурсами.

3.Кількісна оцінка внеску окремих факторів у продуктивність культур за допомогою моделей класифікації/регресії та порівняння їхніх метрик якості.

4.Виявлення стійких патернів та формування сегментів регіонів/культур за агрономічними характеристиками із застосуванням кластеризації та асоціативних правил.

Перелік графічного матеріалу (за потреби)

Дата видачі завдання "01" листопада 2024 р.

Керівник магістерської кваліфікаційної роботи _____ Белла Голуб
(підпис) (ім'я прізвище)

Завдання прийняв до виконання _____ Олександр Пухальський
(підпис) (ім'я прізвище студента)

Календарний план

№ з/п	Назва етапів виконання магістерської кваліфікаційної роботи	Строк виконання етапів магістерської кваліфікаційної роботи	Примітка
1	Видача завдання	01.11.2024	
2	Аналіз предметної області	02.11.2024-22.11.2024	
3	Моделювання предметної області	23.11.2024-25.12.2024	
4	Постановка завдання	26.12.2024-05.01.2025	
5	Проектування системи	06.01.2025-12.02.2025	
6	Розробка системи	13.02.2025-25.03.2025	
7	Аналіз результатів	26.03.2025-15.04.2025	
8	Оформлення записки	16.04.2025-20.09.2025	
9	Проходження нормо контролю	12.11.2025	
10	Перевірка на плагіат	13.11.2025	
11	Попередній захист	24.11.2025-29.11.2025	
12	Захист	05.12.2025-13.12.2025	

Студент _____ Олександр Пухальський

(підпис) (ім'я прізвище)

Керівник магістерської кваліфікаційної роботи _____ Белла Голуб

(підпис) (ім'я прізвище)

РЕФЕРАТ

У роботі розроблено систему, що поєднує технології сховищ даних та інтелектуального аналізу для управлінських рішень у сфері агровиробництва.

Об'єкт дослідження. Процеси вирощування агрокультур у сучасному сільському господарстві.

Предмет дослідження. Система аналітичної обробки аграрних даних на основі сховища даних, OLAP та методів Data Mining для оцінювання й прогнозування врожайності агрокультур.

Мета роботи. Обгрунтувати й перевірити на практиці підхід до поєднання сховища даних, OLAP і Data Mining для отримання пояснюваних моделей врожайності та їх використання у прийнятті управлінських рішень.

Використані методи. Багатовимірний аналіз, класифікація, пошук асоціативних правил, кластеризація, візуалізація результатів.

Вихідні дані. державна агростатистика, кліматичні та грунтові показники, інтегровані у тематичне сховище даних.

Наукова складова. Досліджено поєднання OLAP-кубів із Data Mining для кількісної оцінки внеску факторів (регіон, культура) у варіації врожайності.

Рекомендації щодо впровадження результатів. Використати побудоване сховище та OLAP-моделі для регулярного моніторингу, алгоритму класифікації, асоціативного аналізу і кластеризації для періодичного перегляду карти ризиків.

Ключові результати. Застосування комплексного підходу поглиблює розуміння структури агроданих і створює основу для підвищення ефективності агровиробництва.

Кількість сторінок – 61

Кількість ілюстрацій – 28

Кількість таблиць – 2

Кількість додатків – 1

Кількість джерел – 25

ABSTRACT

The work presents the development of a system that combines data warehouse technologies and data mining methods to support managerial decision-making in agricultural production.

Object of study. Processes of crop cultivation in modern agriculture.

Subject of study. The analytical data processing system based on a data warehouse, OLAP, and Data Mining methods for evaluating and forecasting crop yields.

Purpose of the work. To substantiate and experimentally verify an approach that integrates a data warehouse, OLAP, and Data Mining for building explainable yield models and applying them in decision-making processes.

Methods used. Multidimensional analysis, classification, association rule mining, clustering, and visualization of results.

Input data. State agricultural statistics, climatic and soil indicators integrated into a thematic data warehouse.

Scientific novelty. The combination of OLAP cubes and Data Mining methods has been explored to quantitatively assess the contribution of factors to yield.

Recommendations for implementation. The developed data warehouse and OLAP models can be used for regular monitoring, while classification, association, and clustering algorithms can support periodic review of regional risk and potential maps.

Key results. The integrated approach deepens the understanding of agricultural data structure and forms a foundation for improving the efficiency of agricultural production.

Number of pages – 61

Number of illustrations – 28

Number of tables – 2

Number of appendices – 1

Number of references – 25

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

СД - сховище даних.

AI - Artificial Intelligence

1R - One Rule algorithm

DBMS - Database Management System

OLAP - On-Line Analytical Processing

JSON - JavaScript Object Notation

SQL - Structured Query Language

SSAS - SQL Server Analysis Services

BI - Business Intelligence

SSIS - SQL Server Integration Services

SSRS - SQL Server Reporting Services

PCA - Principal Component Analysis

KPI - Key Performance Indicator

ВСТУП

Актуальність. Цифрова трансформація аграрного сектору супроводжується різким зростанням обсягів і різнорідності даних: від офіційної агростатистики до кліматичних та грунтових показників. Для прийняття обгрунтованих рішень щодо планування посівів, вибору сортів і технологій вирощування потрібні інструменти, що не лише об’єднують дані, а й забезпечують багатовимірний аналіз та виявлення прихованих закономірностей. Інтеграція тематичного сховища даних (СД) з OLAP і методами інтелектуального аналізу даних (Data Mining) створює основу для оцінювання та прогнозування врожайності, побудови карт ризиків і визначення сприятливих комбінацій регіон-грунт-культура. Це підвищує ефективність управлінських рішень у рослинництві та сприяє сталому використанню ресурсів.

Об’єкт дослідження. Процеси вирощування агрокультур у сучасному сільському господарстві.

Предмет дослідження. Інтелектуальна система аналізу параметрів вирощування агрокультур.

Мета роботи. Розробити та застосувати аналітичну систему для оцінювання й прогнозування врожайності агрокультур із використанням технологій Data Mining і OLAP, орієнтовану на підтримку управлінських рішень у рослинництві.

Завдання дослідження:

- Спроєктувати тематичне СД та побудувати вимірні моделі для агроданих.
- Розробити OLAP-куби та набір багатовимірних зрізів/вимірів (регіон, культура, грунт, час).
- Реалізувати Data Mining: класифікація (1R, Наївний Байєс), асоціативні правила (Apriori), кластеризація (K-Means із добором k методом ліктя).

- Оцінити моделі за релевантними метриками (точність, support/confidence/lift, індекс силуету) та виконати порівняння з базовими підходами.
- Провести факторний/варіаційний аналіз внеску регіону, типу грунту, культури у змінність врожайності.
- Створити KPI-панелі для демонстрації процесів в режимі реального часу.
- Забезпечити відтворюваність: зафіксувати версії ПЗ, підготувати інструкції запуску та артефакти (скрипти, конфіги, журнали експериментів).

Методи та засоби. Багатовимірний аналіз (OLAP), класифікація (1R, Наївний Байєс), пошук асоціативних правил (Apriori), кластеризація (K-Means, добір k методом ліктя), візуалізація результатів (у т.ч. PCA). Технологічна база: SQL, SSAS, SSIS/SSRS, Power BI, Python (pandas, scikit-learn, mlxtend).

Джерела даних. Відкриті набори даних: державна агростатистика, грунтові показники, інтегровані у тематичне СД з подальшим завантаженням в аналітичні моделі.

Наукова складова. Досліджено поєднання OLAP-кубів із Data Mining для кількісної оцінки внеску факторів (регіон, тип грунту, культура) у варіації врожайності.

Практична значущість. Результати дозволяють прогнозувати врожайність з урахуванням агрокліматичних умов, ідентифікувати сприятливі поєднання “регіон-грунт-культура”, формувати рекомендації щодо вибору сортів і агротехнологій у регіональному розрізі та підтримувати управлінські рішення за допомогою інтерактивних звітів і KPI-панелей.

Апробація результатів.

1. VII Всеукраїнська науково-практична конференція студентів і аспірантів "Теоретичні та прикладні аспекти розробки комп'ютерних систем `2025" (м. Київ, 2025 р.);

2. II Міжнародна науково-практична конференція "Актуальні питання розвитку науки та техніки в умовах глобалізації" (м. Боярка, 2025 р.);
3. XVI Міжнародна науково-практична конференція молодих вчених "Інформаційні технології: економіка, техніка, освіта" (м. Київ, 2025р.)

Структура роботи Дослідження складається з чотирьох взаємопов’язаних розділів. Розділ 1, в якому сформульовано аналіз об’єкту дослідження. Проведено огляд готових рішень з порівнянням функціональності й обмежень. Сформовано постановку задачі. Розділ 2, в якому подано загальні положення моделювання та UML-опис: діаграма прецедентів, діаграми послідовності, діаграма класів, діаграма розгортання. Окремо наведено концептуальну модель даних і схему “зірка”, що лягає в основу аналітики. Розділ 3, де описано джерела та підготовку даних, проєктування тематичного сховища розгортання OLAP-куба в SSAS, налаштування груп та процесингу. Розділ 4 в якому застосовано методи Data Mining для отримання нових знань: 1-Rule, наївний Байєс, Apriori, K-Means, а також візуальні засоби (KPI) для оперативного аналізу.

1 АНАЛІЗ ПРОЦЕСІВ ВИРОЩУВАННЯ АГРОКУЛЬТУР ЯК ПРЕДМЕТ ДОСЛІДЖЕННЯ

1.1 Опис предметної області

Вирощування агрокультур - це багатокомпонентний процес, що охоплює планування посівів, підготовку грунту, вибір сортів і технологій, сівбу, догляд за посівами, моніторинг стану посівів, збирання врожаю, первинну доробку та зберігання, а також післяврожайний аналіз і планування наступного циклу. Процеси циклічні (річний/сезонний ритм), просторово розподілені (поля, ділянки, кластери господарств), сильно залежні від погодних та грунтово-кліматичних умов і регламентуються як агротехнічними стандартами, так і економічними обмеженнями (ресурсами, бюджетом, логістикою).

Загалом є такі стандартні етапи виробничого циклу (з погляду даних і рішень)

- Передпосівне планування: вибір культур і сортів, розподіл площ за полями/ділянками, формування сівозміни, оцінка ресурсів, прогноз попиту/цін.
- Підготовка грунту: аналіз агрохімії (pH, NPK, органіка), вибір обробітку, норми внесення добрив (базові/змінні карти).
- Сівба: календар і строки, норми висіву, параметри сівалки (глибина, міжряддя).
- Догляд за посівами: моніторинг вологи, температури, шкідників/хвороб, прийняття рішень щодо зрошення, підживлення, обробок, адаптація норм за зонами поля.
- Збирання врожаю: оптимальне вікно збирання, маршрути техніки, фіксація фактичної врожайності (карти врожайності), втрати при збиранні.

- Післяврожайний етап: доробка/зберігання, логістика, реалізація. Аналітика сезону - порівняння план/факт, факторний аналіз, оновлення карт ризиків і потенціалу.

Проблеми та виклики підприємств які можуть бути:

- Фрагментація джерел: розрізнені Excel, паперові акти, окремі системи техніки, метеодані - складно отримати єдину картину.
- Неузгодженість термінології й одиниць: різні імена культур/сортів, різні одиниці виміру, відсутність єдиних ключів.
- Неповнота/шум у первинних даних: пропуски сенсорів, відсутність карт врожайності, ручні виправлення без протоколів.
- Обмежений аналітичний інструментарій: описова звітність без глибинної діагностики, відсутність багатовимірних розрізів та інтелектуального аналізу.
- Ризикова залежність від погоди і нестабільності цін: потрібні сценарії та ймовірнісні оцінки, а не лише середні.
- Локальність практик: успішні кейси на окремих полях важко масштабувати без порівнянної аналітики.
- Повільні цикли зворотного зв’язку: інсайти приходять постфактум (після сезону), тож можливості оперативного коригування втрачаються.

Предметна область вирощування агрокультур характеризується сезонністю, високою варіативністю природних факторів і великою кількістю розрізнених даних. Без спеціалізованих аналітичних інструментів підприємства стикаються з фрагментацією інформації, несумісністю форматів і браком причинно-пояснювальної аналітики.

1.2 Аналіз існуючих програм-аналогів

1.2.1 Agrivi (FMS/BI/Traceability) - екосистема цифрових рішень для агровиробників і харчових компаній: AGRIVI 360 Farm Management, AGRIVI Food / Supply Chain , AI-консультант, інтеграції з метео, IoT та технікою. Наголос на створенні цифрового двійника господарства, плануванні сезону, моніторингу поля, аналітиці та звітності. Тестування системи показані на рис.1.2.1

Рис 1.2.1 Agrivi система

Ключові можливості (релевантні нашому предмету):

— централізований збір польових, метео та сенсорних даних, plug-and-play інтеграції (погода, грунтові сенсори, машини) для моніторингу в реальному часі

— планування/бюджетування сезону, управління роботами, аналітика врожайності та витрат

— супутникові знімки, попередження щодо шкідників/погоди

— модулі трасованості по ланцюгу постачання (для харчових компаній)

Сильні сторони. Широка предметна покритість (від поля до ланцюга постачання), інтеграції з апаратурою/ERP, орієнтація на управлінські рішення.

Обмеження. Аналітика й візуалізації із коробки сильні, але поглиблені ML-конвеєри (моделі з власними метриками/валідаціями) зазвичай потребують зовнішнього сервісу/кастомізації.

1.2.2 Climate FieldView (Bayer) – “завдяки інноваційним інструментам, що використовують системи обробки інформації та/або наукові дані, платформа цифрового землеробства Climate FieldView компанії Bayer дозволяє приймати повсякденні рішення.”[17] Цифрова платформа все-в-одному для фермерів: збір польових даних, моніторинг, аналіз урожайності, створення скриптів (prescriptions) за зонами та цілями урожайності. Є FieldView Drive для збирання даних із техніки, мобільні/веб-додатки, тарифи з передплатою. На рис.1.2.2 представлений приклад збору польових даних системою:

Рис 1.2.2 Збір польових даних

Ключові можливості:

— мапування полів, збирання і історичне завантаження даних, порівняння гібридів/скриптів

— моніторинг прогресу робіт, віддалений перегляд, аналіз урожайності
— створення змінних норм для висіву/внесення добрив на основі зон і цілей

Ціноутворення.

Офіційно пропонуються платні плани (Plus/Premium) .

Приклад: старт від ~$249/рік у США (промо-інформація Bayer/FieldView). Фактичні пакети відрізняються за ринком.

Сильні сторони. Сильний польовий рівень: збір даних, швидкі практичні інструменти, інтеграція з технікою через FieldView Drive.

Обмеження. Платформа орієнтована на фермерський рівень прийняття рішень. Розширені, кастомні ML-моделі з поясненнями і системний what-if для розподілу ресурсів на рівні господарства/регіону поза із коробки, можливе через зовнішню аналітику.

1.2.3 OneSoil - Платформа точного землеробства з безкоштовними базовими інструментами (web+mobile) для моніторингу вегетації, планування робіт та економії ресурсів. PRO-версія додає розширені індекси, VRA-карти тощо. Акцент на супутниковому моніторингу і простоті старту.

Ключові можливості:

— дистанційний моніторинг полів за індексами вегетації, виявлення проблемних зон
— базові інструменти планування. У PRO - нові індекси та карти на основі додаткових даних (електропровідність грунту тощо).

Сильні сторони. Низький бар’єр входу (швидка реєстрація, безкоштовна база), фокус на супутниковій аналітиці як джерелі ознак для прийняття рішень. Підкреслена увага до приватності даних.

Обмеження. Обмежена глибина операційних і бізнес-моделей із коробки. Для системної OLAP/ML-аналітики та управління ресурсами

потрібна інтеграція з нашим сховищем і додатковими даними (агростатистика, грунти, фінанси).

1.2.3 Наукові підходи й релевантні дослідження. Орленко Н.С., Карпич М.К., Коховська І.В. “Особливості сховища даних та оброблення результатів кваліфікаційної експертизи сортів рослин”.

Автори описують практичну побудову сховища даних для результатів випробувань сортів у мережі з 24 пунктів дослідження (Полісся, Лісостеп, Степ), демонструють розвідувальний аналіз, описову статистику й ANOVA/кластеризацію для виявлення маргінальних значень та факторів варіації врожайності.

Ключова ідея - у DW слід зберігати не лише результати, а й контекст (природно-кліматичні умови, прив’язані до фенологічних фаз) і вести історичні «шари» даних для коректних трендів. Рекомендовано інтегрувати зовнішні кліматичні джерела (на кшталт Climate FieldView тощо) і використовувати DW як базу для інтелектуального аналізу (кластерний/дисперсійний аналіз).

Для нашої роботи це прямо обгрунтовує: вибір моделі “зірка” з вимірами часу/регіону/культури/(клімат/грунт), інклюзію кліматичних атрибутів на рівні сезонів/феностадій, практику інкрементального завантаження та історизації, подальше застосування OLAP + Data Mining на узгоджених даних.

З дослідження можна виокремити корисну інформацію “доведено важливість збереження даних про природно-кліматичні умови відповідно до фенологічних стадій розвитку рослин у сховищі даних інформаційної системи; сховище виступає підгрунтям для інтелектуального аналізу результатів експертизи.”[16]

1.3 Постановка завдання

Метою є створення інтелектуальної аналітичної системи для рослинництва, що перетворює розрізнені агродані на пояснювані висновки і керовані дії. Рішення має працювати як єдиний контур: від прийому офіційної агростатистики й довідників до багатовимірних зрізів, моделей виявлення закономірностей та візуальних індикаторів досягнення цілей.

Приймання офіційних наборів (врожайність/площа/виробництво за роками й регіонами), довідників адміністративного поділу, класифікацій культур, за можливості - ознак грунту та клімату. Нормалізація одиниць (ц/га - т/га), уніфікація назв, контроль цілісності ключів.

Тематична модель “зірка” з фактом врожайності та вимірами часу, регіону/ділянки, культури, (опційно грунту, клімату, сорту). На основі цієї схеми - багатовимірні зрізи і узгоджені обчислювані міри.

Моделі знань:

- швидкі однофакторні правила (1R) для первинного виявлення сильних сигналів
- імовірнісна оцінка класу врожайності (Наївний Байєс)
- асоціативні правила (Apriori) для рекомендацій
- сегментація умов (K-Means з підбором K методом ліктя) та інтерпретація через PCA

Аналітичні запити, на які система має відповідати:

- де й за яких поєднань умов ймовірність високої врожайності максимальна, а де підвищений ризик просідання;
- які патерни зустрічаються системно і придатні для масштабування практик;
- які однорідні сегменти (кластери) виділяються за потенціалом і стабільністю, і які управлінські дії доцільні для кожного типового профілю;

- наскільки стійкими в часі є виявлені зв’язки та як вони відбиваються на KPI.

Очікуваним результатом є готова до використання аналітична платформа, яка:

- дає цілісну картину врожайності у вимірах;
- перетворює дані на пояснювані прогнози, правила та сегменти
- формує практичні рекомендації для планування посівів, вибору сортів і технологій
- підтримує моніторинг прогресу через KPI і швидке звернення до причин відхилень

2 ПРОЕКТУВАННЯ СИСТЕМИ АНАЛІЗУ ПРОЦЕСІВ ВИРОЩУВАННЯ АГРОКУЛЬТУР

2.1 Загальні положення моделювання

Моделювання – “це процес створення спеціальних віртуальних або математичних представлень реальних об'єктів, систем або явищ. Це поняття лежить в основі багатьох інноваційних галузей: від науки й медицини до інженерії та розваг.”[18] Загалом це формальне подання суттєвих характеристик системи у вигляді абстракцій (діаграм, специфікацій, обмежень), що дозволяє зрозуміти, спроєктувати та перевірити рішення до його реалізації. Для інформаційно-аналітичних систем у рослинництві моделювання виконує одразу кілька функцій. А саме зменшує складність предметної області через поділ на незалежні погляди (поведінковий, структурний, розгортання), забезпечує трасованість вимог до даних, функцій і показників (KPI) на всіх етапах.

Предметна область характеризується різнорідними джерелами (агростатистика), складними зв’язками час–регіон–культура–грунт і вимогами до пояснюваності результатів (OLAP-зрізи, правила, кластери, прогнози).

Моделі дають можливість:

- узгодити дані до їх завантаження: визначити зерно фактів, ключі вимірів, ієрархії часу/географії, правила конвертації одиниць (ц/га ↔ т/га);
- відокремити OLTP і DWH/OLAP: чітко розмежувати оперативні процеси (введення) і аналітичні (KPI, Data Mining);
- задати пояснюваність: від прецедентів користувача і послідовностей викликів до класів моделі та фізичних вузлів розгортання.

“UML (Unified Modeling Language) — уніфікована мова моделювання, що використовується розробниками програмного забезпечення для візуалізації

процесів та роботи систем.”[25] Це стандарт нотацій для опису ПЗ, який забезпечує зрозумілу, інструментально підтримувану і узгоджену документацію. Він корисний тим, що підтримує кілька типів діаграм для різних кутів зору (поведінка, структура, архітектура).

UML не замінює предметно-орієнтованих нотацій (напр. ER-діаграм чи BPMN), але в нашій роботі виступає основою для узгодження поведінки та структури на рівні системи, тоді як для даних ми паралельно використовуємо схему “зірка”.

2.2 Діаграма прецедентів

Як зазначено в [8] “Діаграма прецедентів використовує 2 основних елементи: Actor (учасник) - множина логічно пов'язаних ролей, виконуваних при взаємодії з прецедентами або сутностями (система, підсистема або клас). Use case (прецедент) - опис окремого аспекту поведінки системи з точки зору користувача.’’

Актори й наміри.

- Фермер - реєструє нові ділянки, веде каталоги культур/насіння, вводить дані про врожайність, переглядає оперативну інформацію.
- Аналітик - завантажує/верифікує дані, виконує OLAP-аналіз, запускає моделі, формує прогнози, готує звіти та рекомендації.
- Власник - оглядає зведені панелі та звіти, проводить порівняльний аналіз, приймає управлінські рішення.

Основні прецеденти.

- Реєстрація нової ділянки (фермер) - створення Area із прив’язкою до Sowing, вказання грунту/клімату.
- Ведення каталогів культур і насіння (фермер/аналітик) - наповнення Culture, Seed.

- Збір даних про врожайність (фермер) - фіксація факту з прив’язкою до ділянки й культури.
- Перегляд даних та аналіз тенденцій (аналітик) - OLAP-зрізи, KPI.
- Прогнозування врожайності (аналітик) - запуск моделей, перегляд інтервалів довіри.
- Виявлення зон з низькою ефективністю та оптимізація вибору насіння (аналітик) - правила Apriori, сегменти K-Means.
- Порівняльний аналіз і прийняття рішень (власник).

На схемі (рис. 2.2) ці прецеденти зв’язані з відповідними акторами.

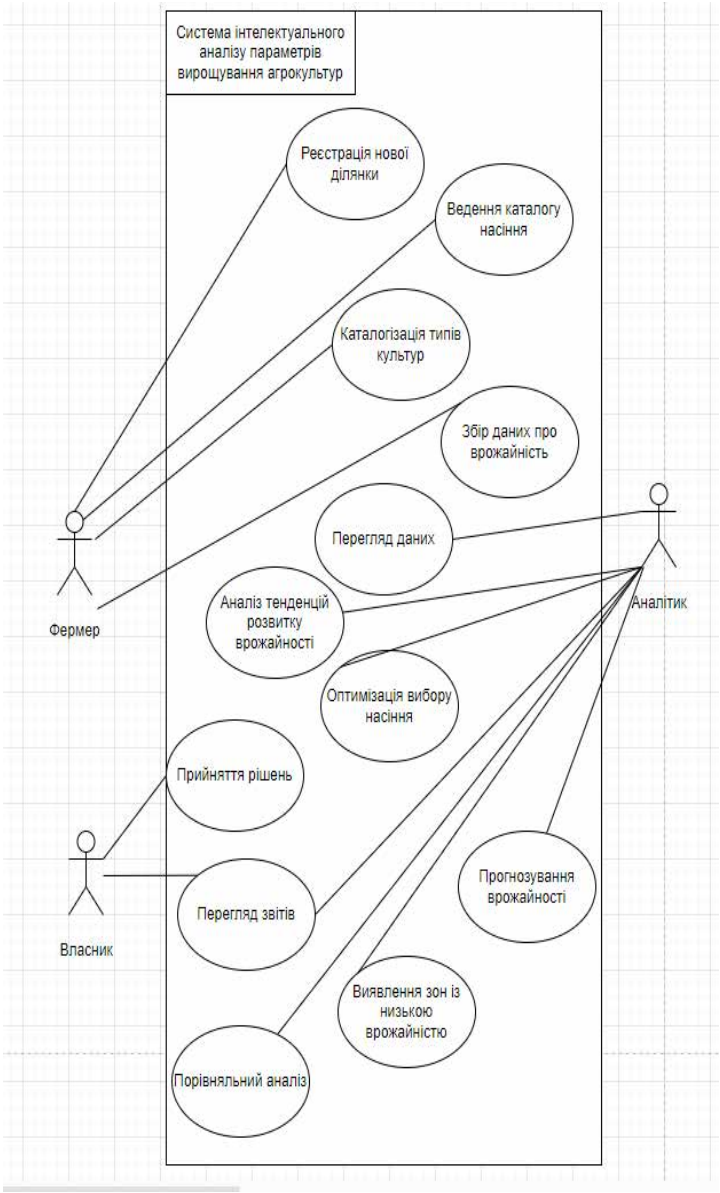


Рис 2.2 Діаграма прецедентів

2.3 Діаграма послідовності

Як зазначено в [5] “Діаграма послідовності (Sequence Diagram) - використовуються для моделювання логіки сценаріїв використання, показуючи інформацію, що передається між об'єктами в системі під час виконання сценарію.” З позицій моделювання, діаграма послідовності відображає часову впорядкованість взаємодій між учасниками сценарію та уточнює порядок викликів: хто ініціює дію, які параметри передаються, у якій послідовності виконуються операції та які результати повертаються. На діаграмі фіксуються життєві лінії об’єктів, типи повідомлень, а також комбіновані фрагменти для альтернатив і циклів.

Додавання культури в БД

Діаграма послідовності для додавання культури в БД показано на рис 2.3.1. Об’єкти: User, Interface_culture (UI-форма), Controller_culture (серверний контролер), Culture (модуль/DAO збереження).

Послідовність:

- Користувач ініціює запит на створення культури, інтерфейс повертає форму для заповнення.
- Після введення даних UI передає їх у Controller_culture.
- Контролер виконує валідацію й передає дані в Culture для збереження.
- Якщо дані коректні - культура додається, формується підтвердження й повертається користувачу.
- Якщо дані некоректні - користувач отримує повідомлення про помилку з переліком полів, що не пройшли перевірку.

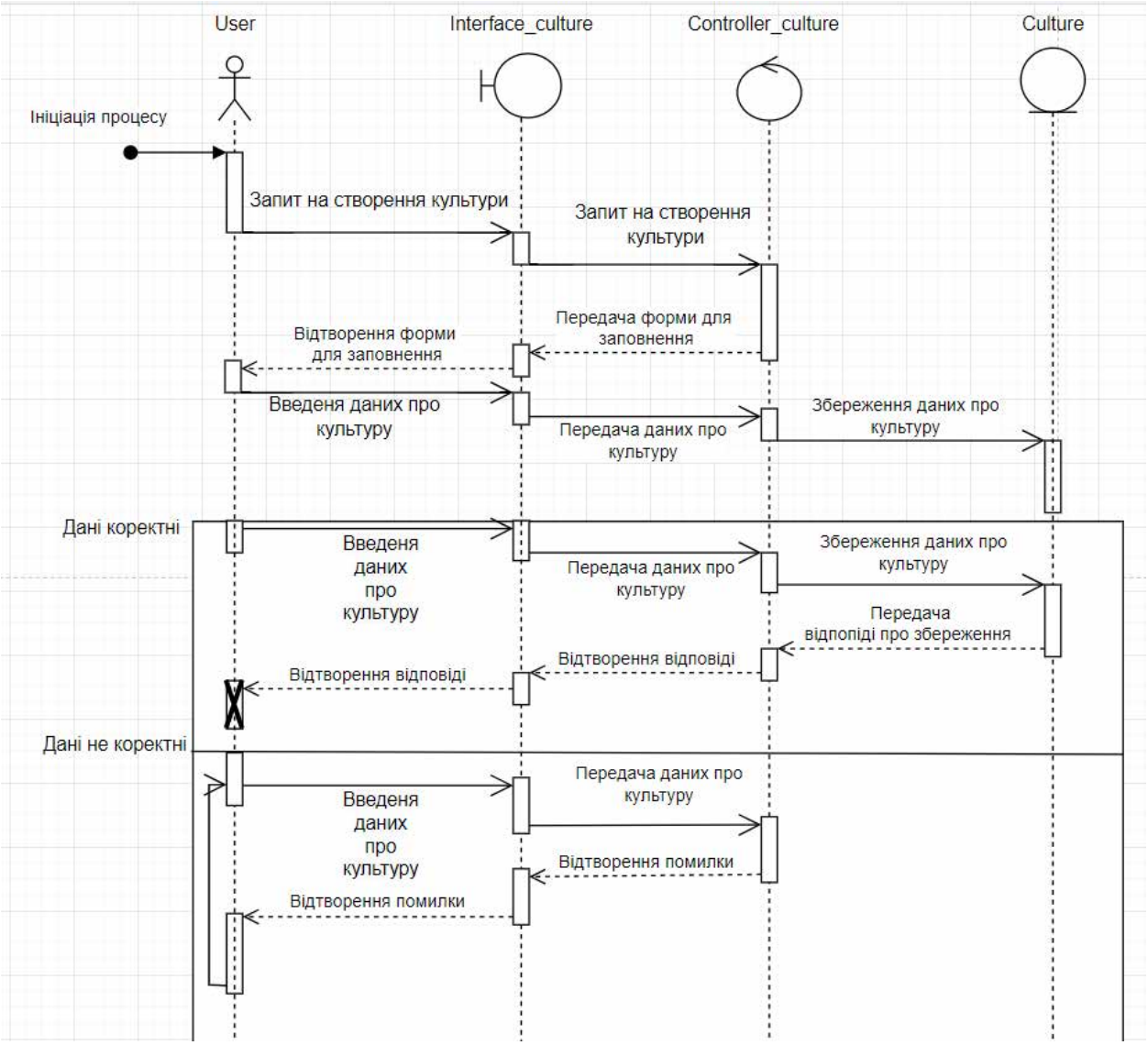


Рис 2.3.1 Діаграма послідовності “Додавання культури в БД”

Додавання посіву в БД

Діаграма послідовності для додавання посіву в БД показано на рис 2.3.2. Об’єкти: User, Interface_sowing, Controller_sowing, Sowing.

Послідовність (узагальнено):

- Користувач подає запит на створення посіву, інтерфейс повертає форму (назва, опис, кількість ділянок, прив’язки до культури/насіння).
- Після введення даних контролер перевіряє повноту/формати і передає їх у модуль Sowing.

- За коректних даних запис зберігається, користувач отримує підтвердження. За некоректних - формується повідомлення про помилку з підсвіченими полями.

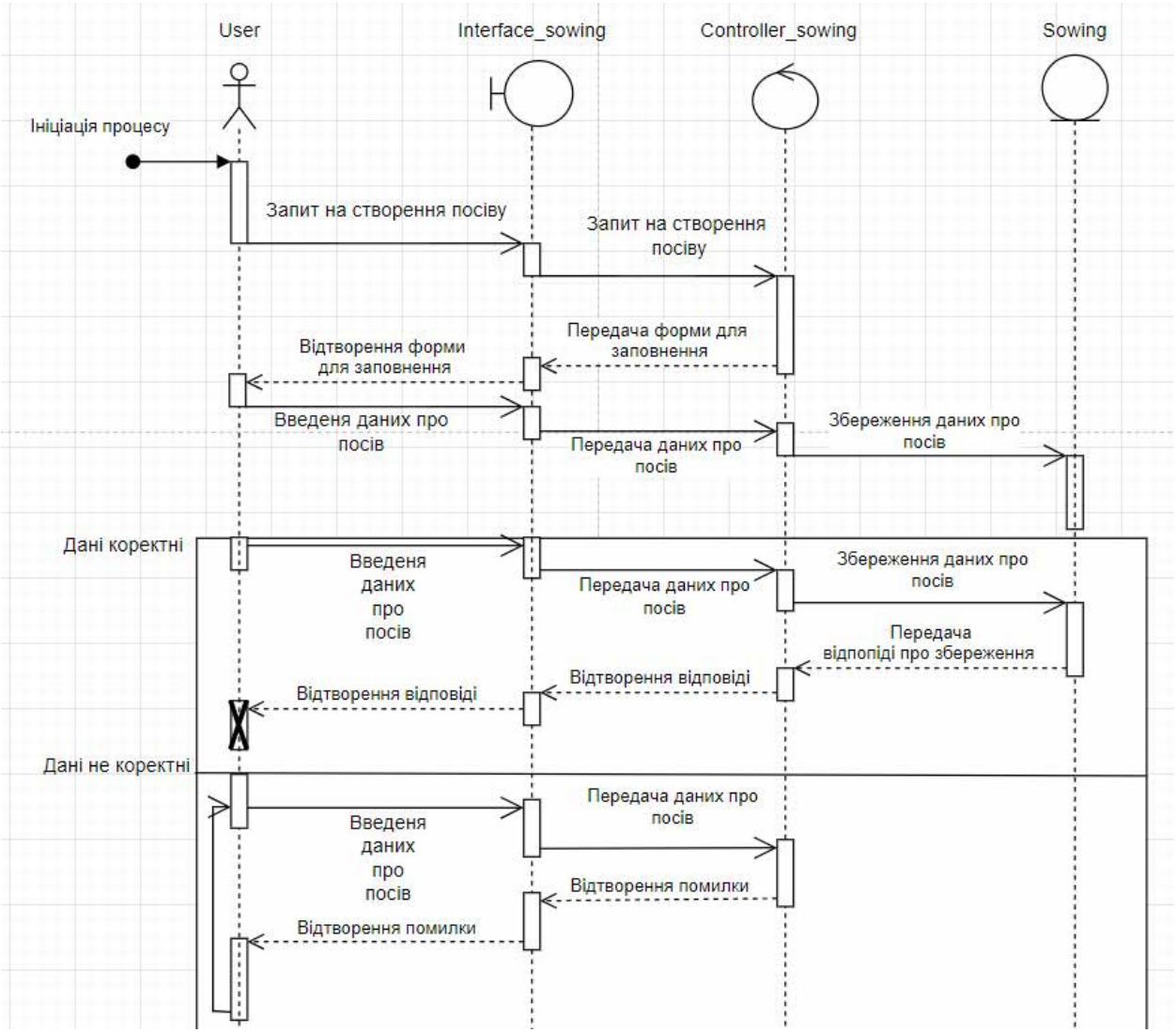


Рис 2.3.2 Діаграма послідовності “Додавання посіву до БД”

2.4 Діаграма класів

“Діаграма класів UML це візуальна нотація, яка використовується для побудови та візуалізації об'єктно-орієнтованих систем. Діаграма класів в уніфікованій мові моделювання — це статична структурна діаграма, що

демонструє властивості системи, класи, операції та зв’язки між об’єктами для опису структури системи.”[24]

Діаграма класів відбиває структурну основу доменної моделі та головні залежності.

Класи та ключові атрибути:

- Account: id_acount, login, hash, ip.
  Відповідальність: автентифікація/авторизація, логування дій.
- Culture: id_culture, name_culture.
  Відповідальність: довідник культур.
- Seed: id_seed, name_seed (+ FK на Culture).
  Відповідальність: сорти/рядки насіння.
- Sowing: id_sowing, name_sowing, description_sowing, number_of_plots_area (+ FK на Culture, Seed, Account).
  Відповідальність: опис технології посіву.
- Area: id_area, number_of_lines_area (+ FK на Sowing).
  Відповідальність: ділянка, грунт/клімат.
- Line: id_line (+ FK на Area).
  Відповідальність: деталізація технологічних операцій у межах ділянки.
- SeedLine: (PK id_seed,id_line), id_area.
  Відповідальність: застосування конкретного насіння на конкретному рядку/ділянці.
- Result: id_result, name_result, data_result (+ FK на Culture, Account).
  Відповідальність: зберігання результатів вимірів/обчислень (у т.ч. агреговані значення врожайності).
  На діаграмі (рис. 2.4) відобразити класи, атрибути, асоціації.

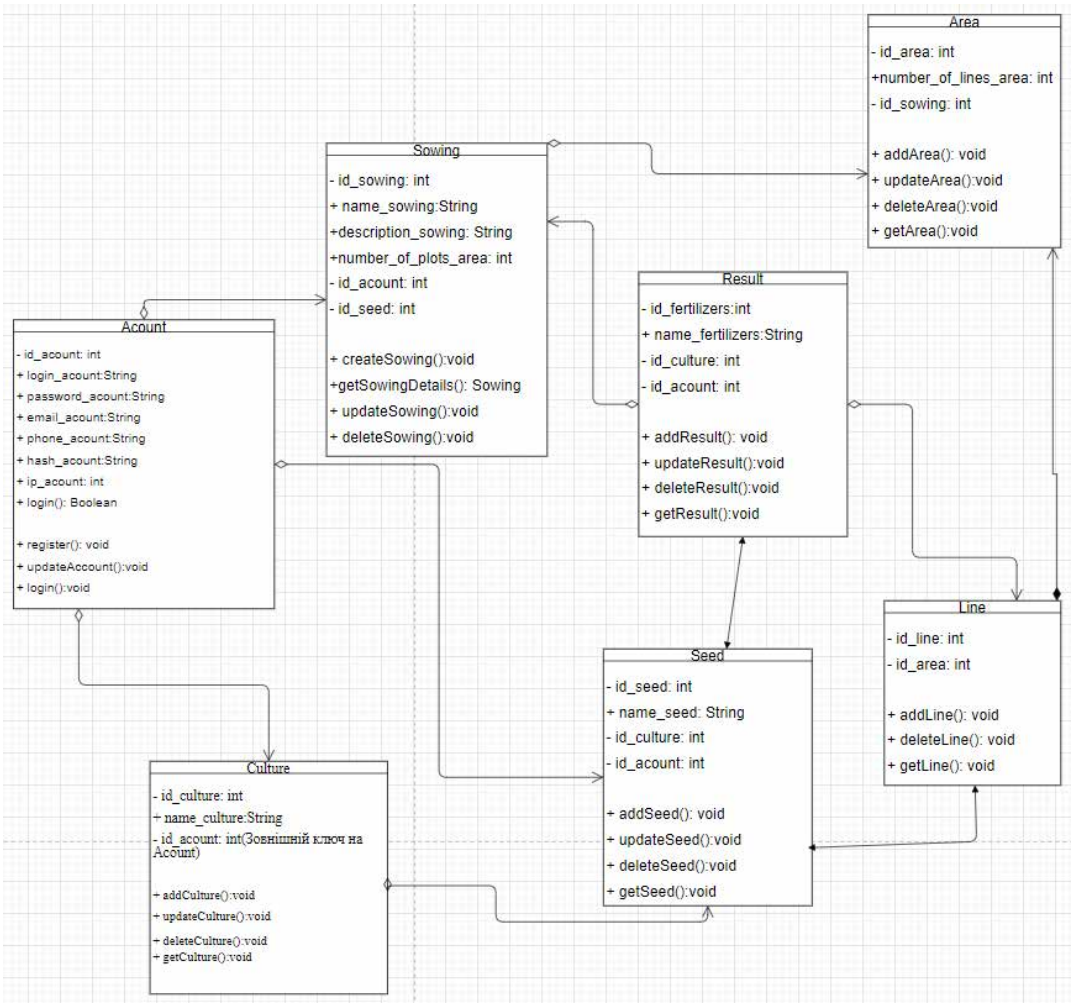


Рис 2.4 Діаграма класів

У діаграмі реалізовано зв’язки, які відображають природну логіку предметної області. Від користувача до культури, від культури до насіння, від посіву до ділянки й рядка.

Така структура моделі робить систему масштабованою й зрозумілою для подальшого розвитку. Вона забезпечує наочне відображення взаємозв’язків між компонентами і полегшує їхню реалізацію в коді.

Отже, діаграма класів на рис. 2.4 демонструє цілісну структуру програмної системи, у якій кожен клас виконує свою роль і пов’язаний з іншими логічними зв’язками. Це дозволяє ефективно реалізувати процеси збереження, обробки та аналізу даних у системі.

3 РОЗРОБКА СИСТЕМИ

3.1 Аналітична архітектура: сховище даних, OLAP та Data Mining

Управління рослинництвом дедалі більше спирається на дані: щосезону накопичуються статистичні звіти, погодні зміни, карти грунтів, виробничі журнали. Проте для прийняття рішень важлива не сама наявність масивів, а здатність швидко перетворювати їх на узгоджені показники, порівнювані зрізи та пояснювані висновки. Саме тому архітектура рішення має будуватися не навколо однієї таблиці, а навколо зв'язки взаємодоповнюваних компонентів: сховища даних (DWH) як єдиного джерела, OLAP як механізму багатовимірного аналізу та Data Mining як інструментів виявлення закономірностей і прогнозування. Далі узгодимо ролі кожного з цих шарів і їхню взаємодію.

Сховище даних (DWH): основа узгодженості - тематично організована, інтегрована база, призначена для аналітики. "Сховище даних працює як центральне сховище, куди надходить інформація з одного або кількох джерел даних."[23]

OLAP: багатовимірний погляд на показники презентує дані DWH у вигляді багатовимірної моделі (куба). "У центрі OLAP-куба — таблиця, що містить ключові факти, за якими роблять запити."[22] Де Факт містить міри (Yield, Production, Area) з прив'язкою до ключів вимірів. Виміри та ієрархії дозволяють переходити між рівнями деталізації. Передобчислені агрегації та індекси забезпечують швидкі відповіді на slice/dice, drill-down/roll-up, pivot, drill-through, top-N. OLAP виступає аналітичним ядром для оперативних порівнянь, трендів і єдиних KPI, гарантуючи однакові формули показників у всіх звітах.

Data Mining: "це аналітичний процес, який досліджує великі обсяги даних з різних точок зору та узагальнює їх таким чином, щоб виявити корисні

закономірності та взаємозв'язки.”[21] Від описових зрізів до знань доповнює OLAP там, де потрібні причинно-пояснювальні та прогностичні висновки:

- Класифікація: 1R (інтерпретовані однофакторні правила) та Наївний Байєс (ймовірності класу High/Low для поєднань регіон–культура–грунт).
- Асоціативний аналіз (Apriori): правила з метриками support/confidence/lift для рекомендацій розміщення культур.
- Кластеризація (K-Means): виділення сегментів умов/регіонів, добір K методом ліктя і візуальна інтерпретація через PCA.

Такі моделі відповідають на запитання чому і з якою ймовірністю, підсилюючи описову аналітику OLAP.

Таким чином, поєднання DWH + OLAP + Data Mining формує повний і керований цикл роботи з агроданими: побачити (OLAP), пояснити та спрогнозувати (Data Mining), закріпити рішення на узгоджених даних (DWH).

3.2 Діаграма розгортання

Як зазначено в [11] “Діаграма розгортання - це тип діаграми, який визначає фізичне обладнання, на якому працюватиме програмна система.”

На рис. 3.2 подано фізичну структуру розміщення компонентів системи. Діаграма демонструє вузли, у яких виконуються прикладні сервіси й зберігаються дані, а також канали обміну між ними. Архітектура побудована з урахуванням розділення транзакційних та аналітичних навантажень, відтворюваності експериментів і безпечного доступу користувачів.

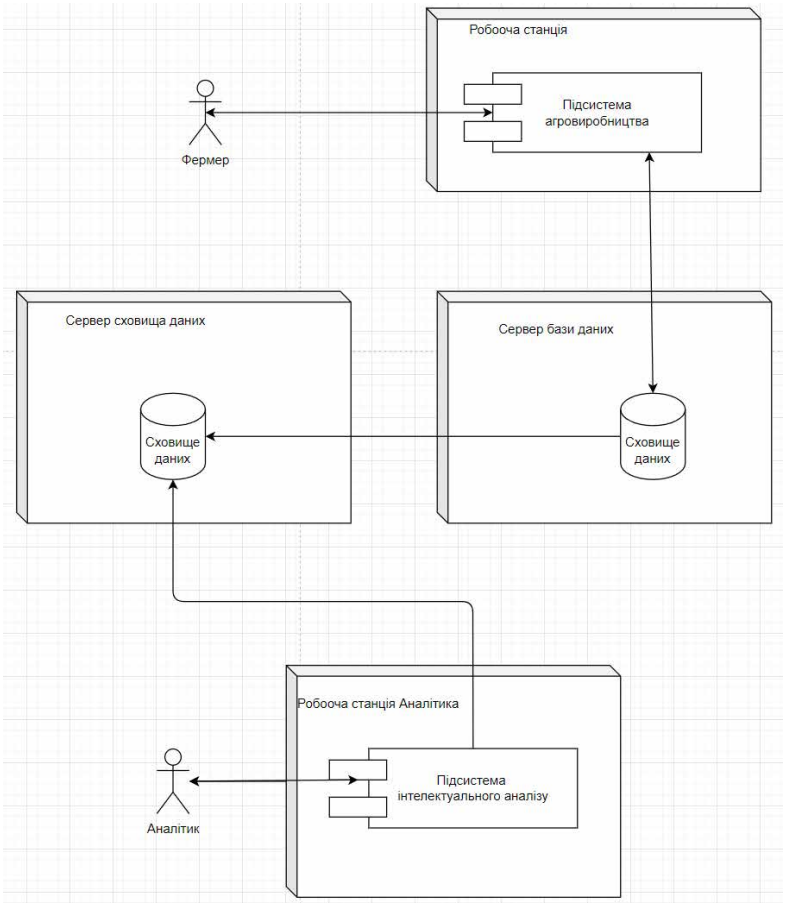


Рис 3.2 Архітектура системи

Робоча станція користувача

- Призначення: введення даних (ділянки, культури, врожайність), перегляд інформації та звітів.
- ПЗ: веб-браузер.
- З’єднання: HTTPS.

Сервер бази даних

- Призначення: транзакційне зберігання первинних записів (облікові записи, посіви, ділянки, культури/насіння, результати).
- ПЗ: СУБД MS SQL Server.

Сервер сховища даних

- Призначення: аналітичне зберігання у моделі “зірка” (Fact_Yield Dim_Area, Dim_Sowing, Dim_Culture, DateDim).

- Зв’язки: отримує дані з OLTP, надає їх OLAP для роботи з кубами.

Робоча станція аналітика

- Призначення: побудова OLAP-зрізів, запуск інтерпретованих моделей (1R, Naive Bayes, Apriori, K-Means), підготовка прогнозів і рекомендацій.

3.3 Джерела даних

У роботі як єдине офіційне джерело фактів для побудови аналітики використано статистику Державної служби статистики України (ukrstat.gov.ua). Базовим матеріалом слугували Excel-файли звітних форм:

- № 29-сг “Звіт про збирання врожаю сільськогосподарських культур” (щомісячна форма)
- № 37-сг “Звіт про збирання врожаю сільськогосподарських культур” (річна зведена форма)

Для забезпечення порівнюваності у часі були відібрані зрізи на 1 грудня для 2021-2024 років. Із цих форм витягнуто три ключові показники, що утворюють золотий трикутник агростатистики:

1. Площа зібраних зернових і зернобобових культур (тис. га),
2. Обсяг виробництва (тис. ц),
3. Урожайність (ц/га).

Окрім фактичних значень, дані збагачено метаданими про адміністративний поділ (область) для коректної геоприв’язки та агрегацій. Усі первинні Excel-файли збережено без змін у вигляді холодного архіву для аудиту та відтворюваності.

Структура полів і одиниці вимірювання

- Area - площа зібраних культур, тис. га.
- Production - валовий збір, тис. ц (еквівалентно тис. 100 кг).
- Yield - ц/га (100 кг з 1 га).

Для аналітики в OLAP усі значення нормалізуються до узгоджених одиниць: Area = га (множення на 1000), Production = т (множення на 10), Yield залишається ц/га або переводиться у т/га (ділення на 10) залежно від обраної політики показників.

До кожного запису додано:

- RegionName -назва області
- DateKey - ключ періоду (рік,)
- CropName -культура

Ці метадані використовуються як виміри в OLAP і дозволяють будувати ієрархії область - район і група - культура.

Отримання та первинна технічна обробка відбувалась таким чином:

1. Завантаження оригінальних файлів із ukrstat.gov.ua за кожний рік.
2. Фіксація контрольної дати (01.12) для 29-сг і відбір підсумкового рядка для 37-сг.
3. Уніфікація форматів Excel (можливі різні аркуші/макети між роками): явне зазначення аркуша, діапазонів, літер роздільників, кодування.
4. Стандартизація заголовків колонок.
5. Збереження без змін оригіналів у архів(первинний шар).

Контроль якості та валідації

Синтаксичний рівень:

- цілісність рядків (відсутність порожніх ключових полів - регіон, культура, дата)
- коректний тип даних (числові поля без дефісів/рядкових артефактів)
- нормалізація пробілів, відсікання службових символів

Семантичний рівень:

- невід’ємність Area, Production, Yield
- повнота покриття: частка пропусків у полях має бути < 5% (або маркуються як Unknown і не враховуються)

3.4 Опис сховища даних

“Схема “зірка” – це широко використовуваний дизайн бази даних у сховищах даних, де транзакційні дані витягуються, трансформуються і завантажуються в схеми, що складаються з центральної таблиці фактів, оточеної узгодженими таблицями розмірностей.”[20]

На рис. 3.4 наведено модель сховища даних “зірка”, яка вже зібрана у SQL Server: у центрі – таблиця фактів Fact_Yield, довкола таблиці вимірів DateDim, Dim_Area, Dim_Sowing, Dim_Seed, Dim_Culture.

Для кожного ланцюга вимірів ділянка * насіння (сорт) * дата був визначений відповідний факт yield (урожайність). Такі виміри дозволяють робити зрізи не лише за культурами, а й за сортами в контексті ділянки й технології посіву.

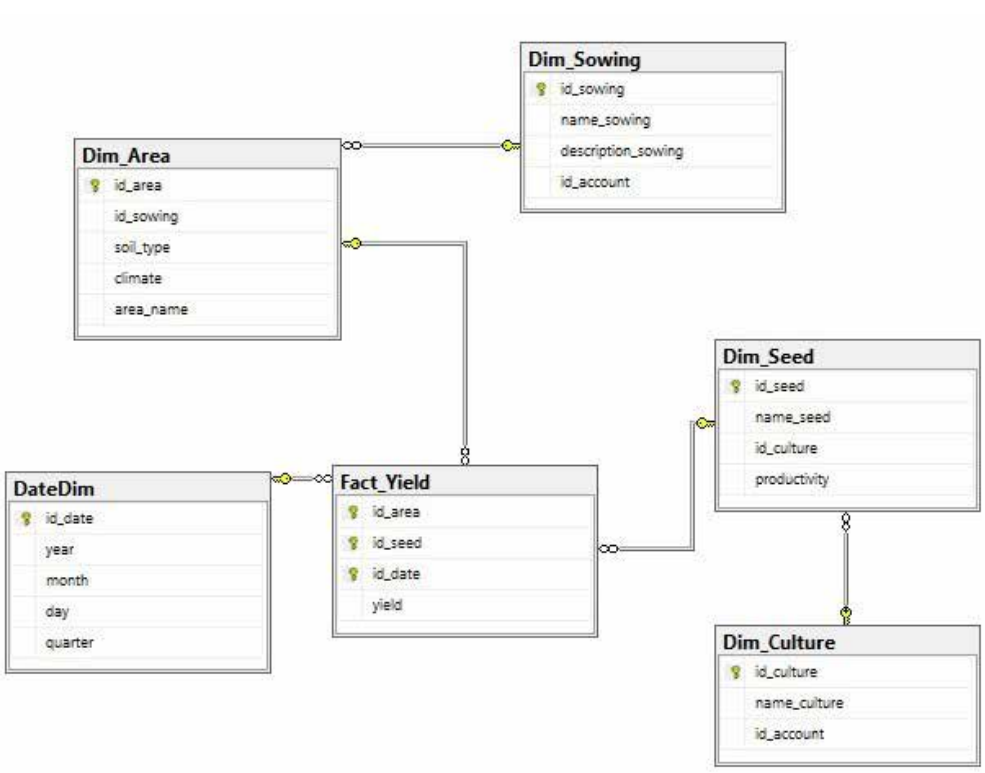


Рис 3.4 Сховище даних

Склад і призначення таблиць

- Fact_Yield (id_area, id_seed, id_date, yield)
  Фіксує врожайність у пункті ділянка-насіння-дата. Композитний PK гарантує унікальність запису на рік.
- DateDim (id_date, year, quarter, month, day)
  Календарний вимір із типовими ієрархіями Year-Quarter-Month-Day.
- Dim_Area (id_area, id_sowing, soil_type, climate, area_name)
  Просторово-агротехнічний вимір: назва ділянки, тип грунту, кліматична ознака, а також зв’язок із технологією посіву (id_sowing).
- Dim_Sowing (id_sowing, name_sowing, description_sowing, id_account)
  Довідник технологій посіву (схема, опис).
- Dim_Culture (id_culture, name_culture) та Dim_Seed (id_seed, name_seed, id_culture, productivity)
  Довідники культур і сортів.

3.5 Розгортання OLAP-куба (SSAS)

OLAP куб було розгорнуто на SQL Server Analysis Services поверх підготовленого DWH (схема “зірка” з Fact_Yield, DateDim, Dim_Area, Dim_Sowing, Dim_Culture, Dim_Seed). Створено єдине підключення до бази даних і сформовано Data Source View, який віддзеркалює логічні зв’язки між таблицями. На рис. 3.5.1 наведено екран майстра DSV та підтвердження створеного підключення.

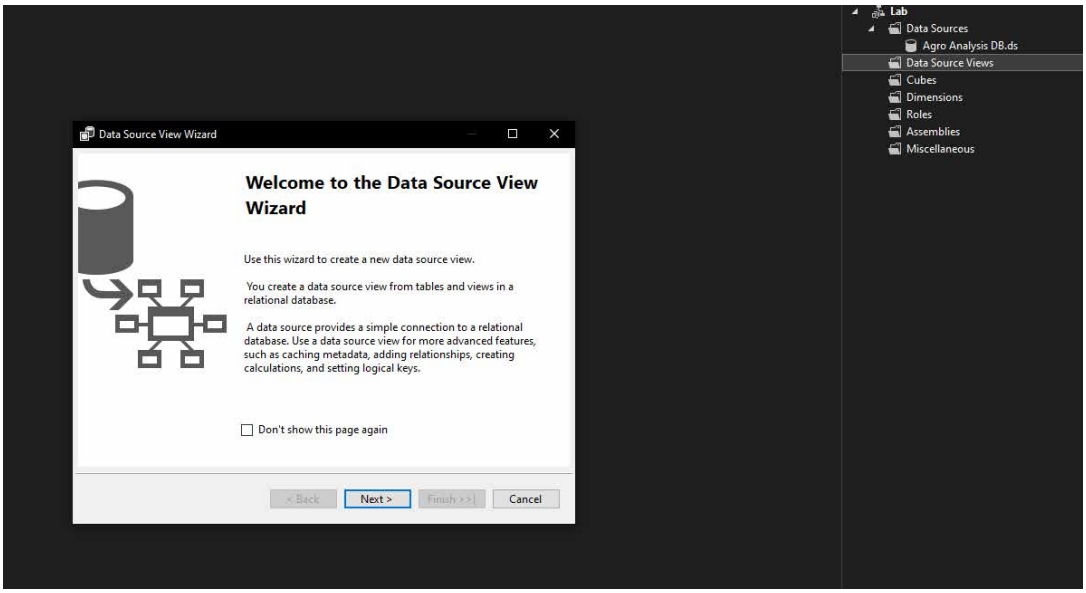


Рис 3.5.1 Майстер Data Source View

Розгорнуту ілюстрацію задіяних таблиць наведено на рис. 3.5.2. Показано повний перелік таблиць, що беруть участь у завантаженні.

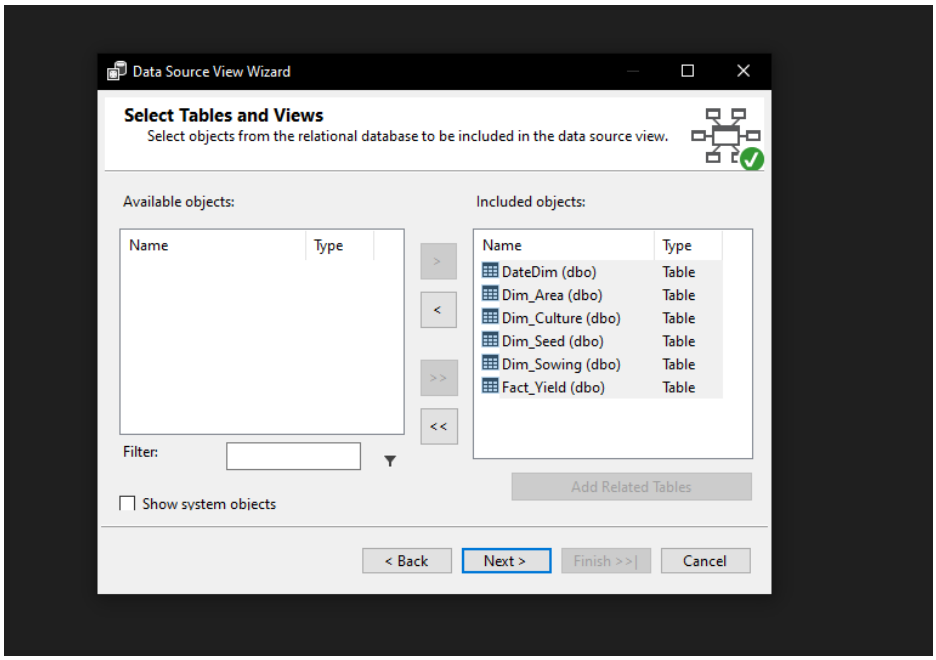


Рис 3.5.2 Перелік включених таблиць

На рис. 3.5.3 зображено фінальне зібране представлення “Agro Analysis DB”. У підсумку, на діаграмі чітко видно факт врожайності та ключі на виміри.

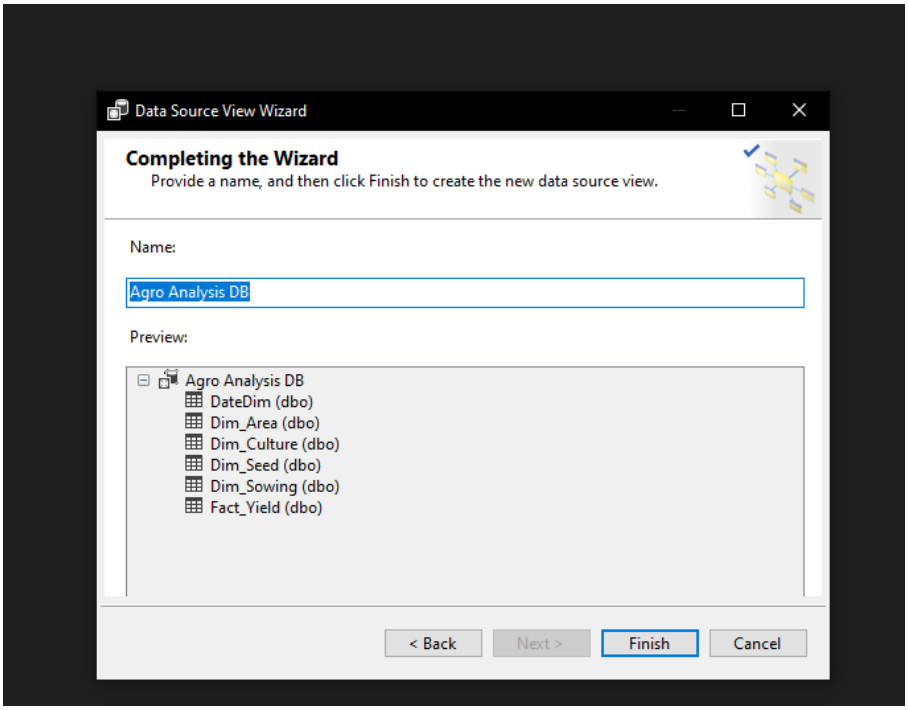


Рис 3.5.3 Підсумок майстра DSV

На базі DSV побудовано куб YieldCube із однією групою мір Fact_Yield та п’ятьма вимірами: Date, Area, Sowing, Culture, Seed. Під час конструювання куба використано існуючі таблиці джерела (див. рис. 3.5.4).

Головною мірою є Yield (урожайність), яка агрегується як середнє у більшості зрізів і як сума у допоміжних агрегаціях. Для часової навігації реалізовано повну ієрархію Year - Quarter - Month – Day. Для предметної області - ієрархію Culture - Seed, а також атрибути soil_type, climate, area_name у вимірі Area.

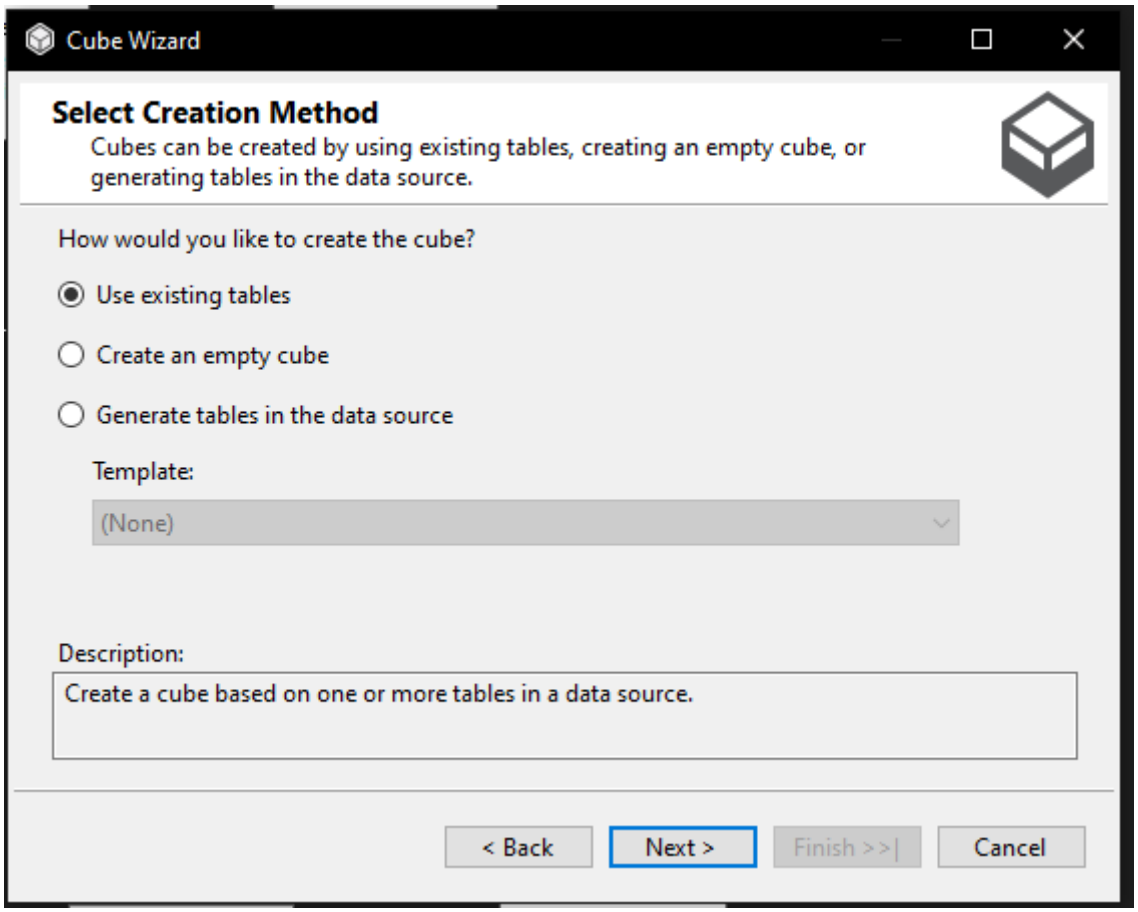


Рис 3.5.4 Вибір підходу “Use existing tables” під час створення куба

На рис. 3.5.5 подано робочу панель проекту з переліком створених вимірів і груп мір, праворуч візуалізацію DSV із відображенням ключових атрибутів, якими користується куб. У структурі джерела даних є взаємозв’язки між таблицями вимірів і таблицею фактів, що забезпечують правильне агрегування показників у процесі аналітичних обчислень. Виміри (Dim_Culture, Dim_Seed, Dim_Sowing, Dim_Area, DateDim) формують ієрархічну модель для деталізації аналітики за культурами, насінням, ділянками, типами грунту й періодами часу. Завдяки цьому куб дає змогу швидко здійснювати зрізи даних, виконувати порівняльний аналіз і формувати аналітичні звіти за різними параметрами агровиробництва.

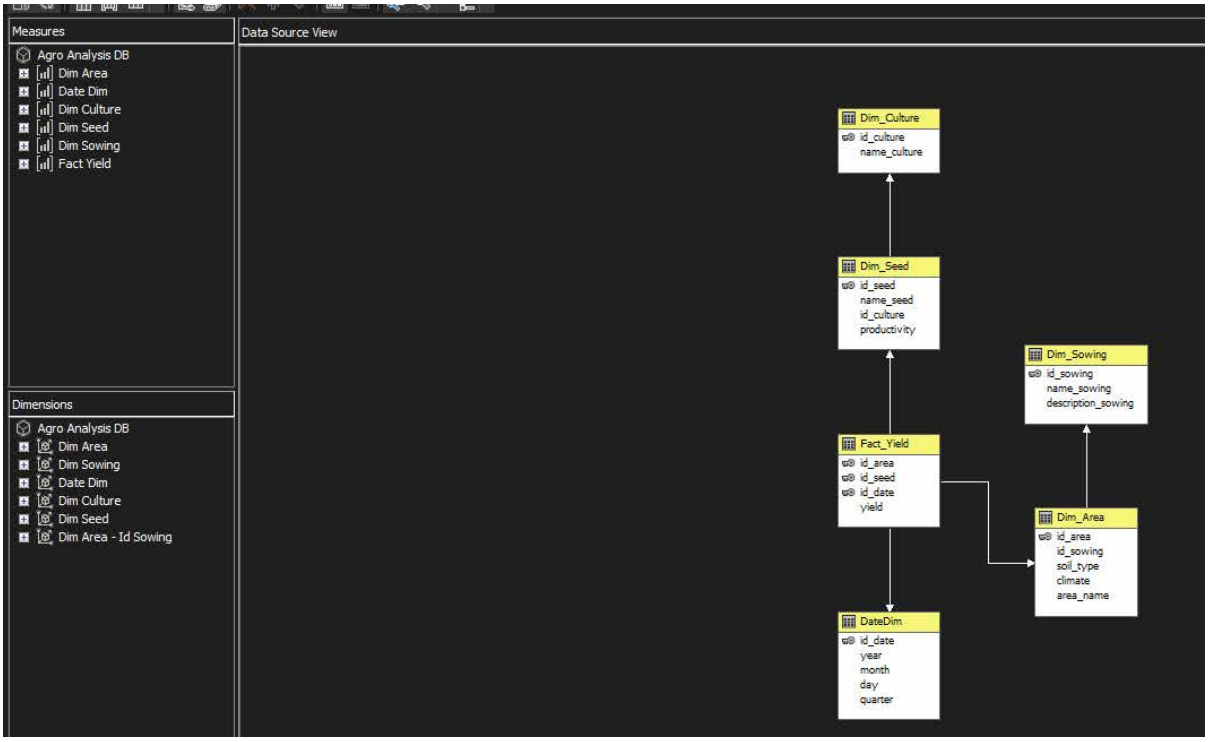


Рис 3.5.5 Вигляд проекту SSAS

Куб налаштовано на оновлення частинами: для групи мір Fact_Yield створені групи за роками, що забезпечує короткий час процесингу та стабільний відгук запитів при зростанні обсягів даних.

3.6 Наповнення кубу даними

“ETL є абревіатурою: Extract, Transform, Load – Витягнення, Трансформація, Завантаження. Це процес, який використовується для переміщення даних з одного або декількох джерел до цільової системи, яка зазвичай є сховищем даних або озером даних (Data Lake).”[19]

ETL-процес реалізовано в SSIS як один пакет, що послідовно заповнює виміри та факт і після цього ініціює процесинг кубу. Логіка побудована за принципом, щоб гарантувати цілісність ключів і коректне агрегування.

Загальна схема пакета. На рис. 3.6.1 подано керуючий потік (Control Flow): два кроки наповнення вимірів і завершальний крок внесення фактів

врожайності. Логічні залежності (зелені стрілки) забезпечують запуск наступного етапу лише після успішного завершення попереднього.

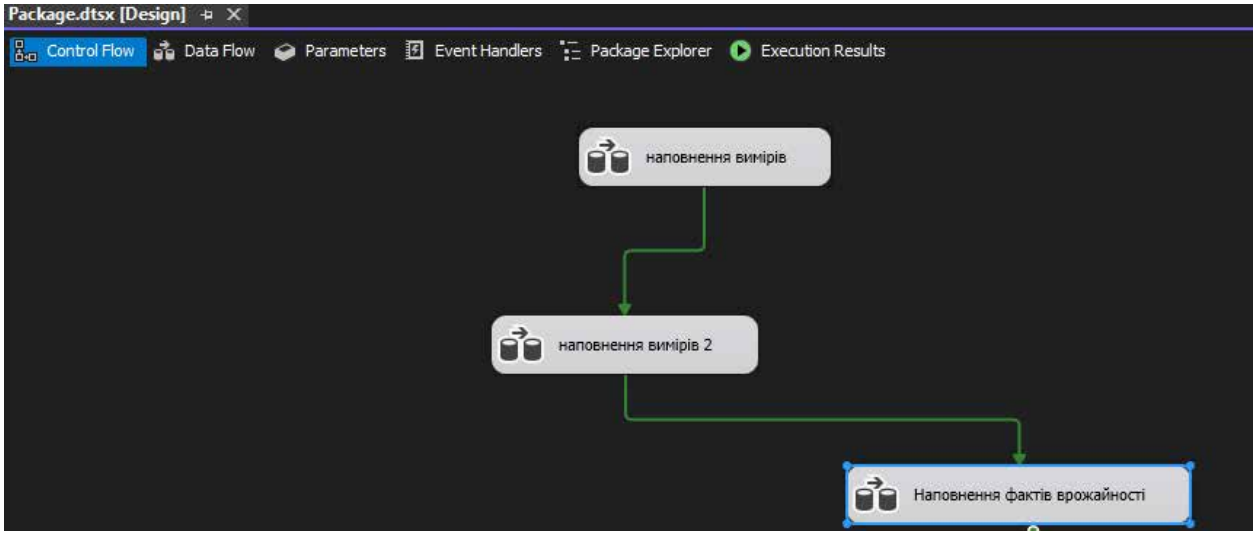


Рис 3.6.1 Керуючий потік пакета SSIS

Потоки даних для вимірів. На рис. 3.6.2 та рис. 3.6.3 показано типові Data Flow для вимірів:

- джерела OLE DB Source читають оперативні таблиці (Sowing, Culture, Area, Seed або підготовлені подання)
- призначення OLE DB Destination записують у відповідні таблиці DWH (Dim_Sowing, Dim_Culture, Dim_Area, Dim_Seed)

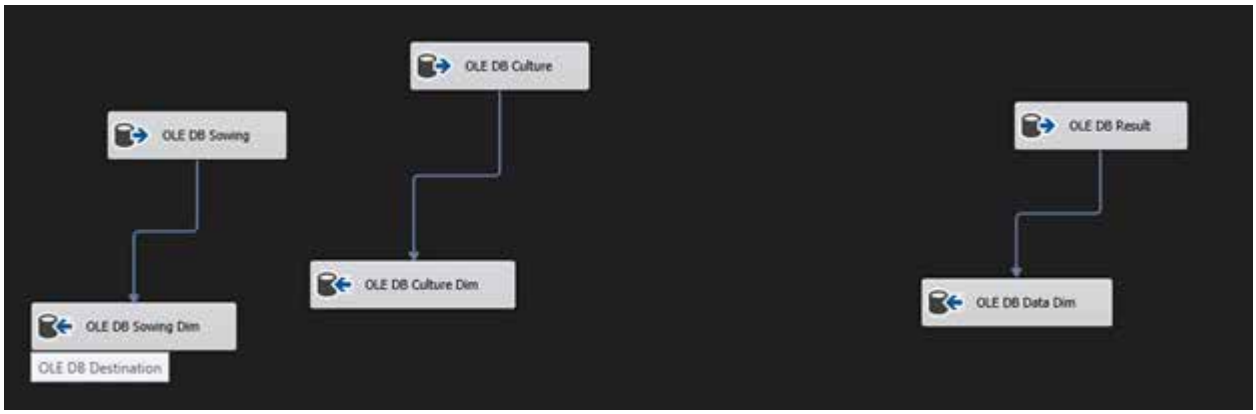


Рисунок 3.6.2 Потік “Наповнення вимірів”

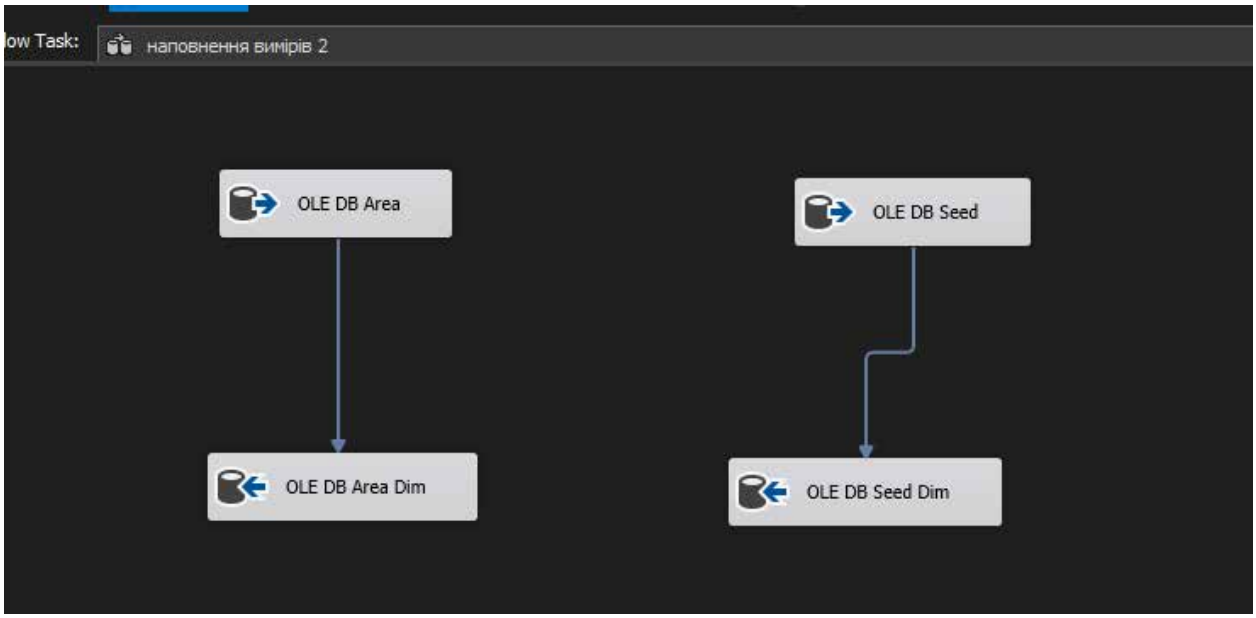


Рисунок 3.6.3 Потік “Наповнення вимірів 2”

Формування календаря. Для DateDim використано SQL-запит, що витягує унікальні комбінації year / month / day / quarter з оперативної таблиці результатів (рис. 3.6.4). Це забезпечує узгодження часової шкали з реально наявними періодами у фактах.

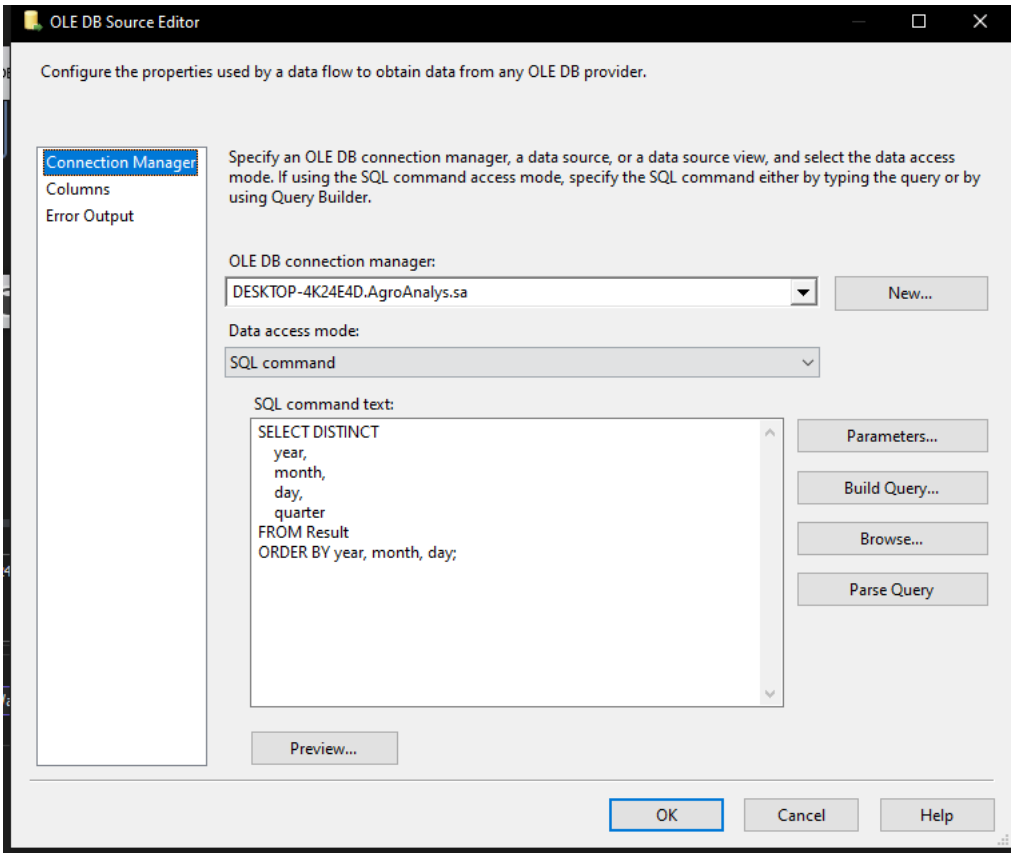


Рисунок 3.6.4 Налаштування джерела для DateDim

Наповнення фактів. Завершальний потік (див. рис. 3.6.5) з’єднує джерело Result із довідниками та DateDim, формуючи зерно ділянка * насіння * дата. Текст запиту із явними JOIN показано на рис. 3.6.6. Він відразу повертає ключі потрібних вимірів та міру yield.

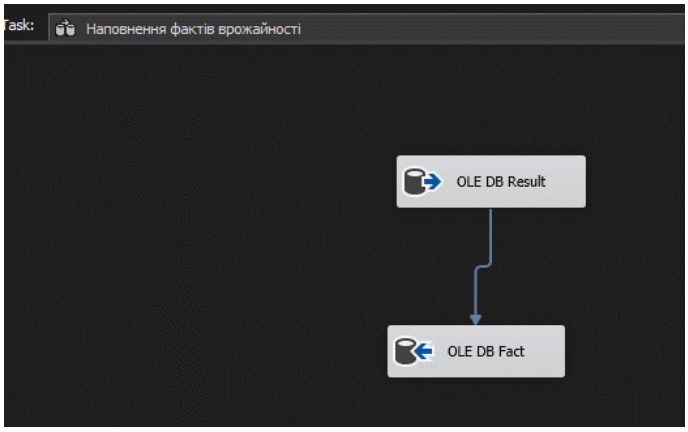


Рисунок 3.6.5 Потік “Наповнення фактів врожайності”

ключі між Result, Area, Seed, Culture, Account і DateDim (зіставлення за year і month).

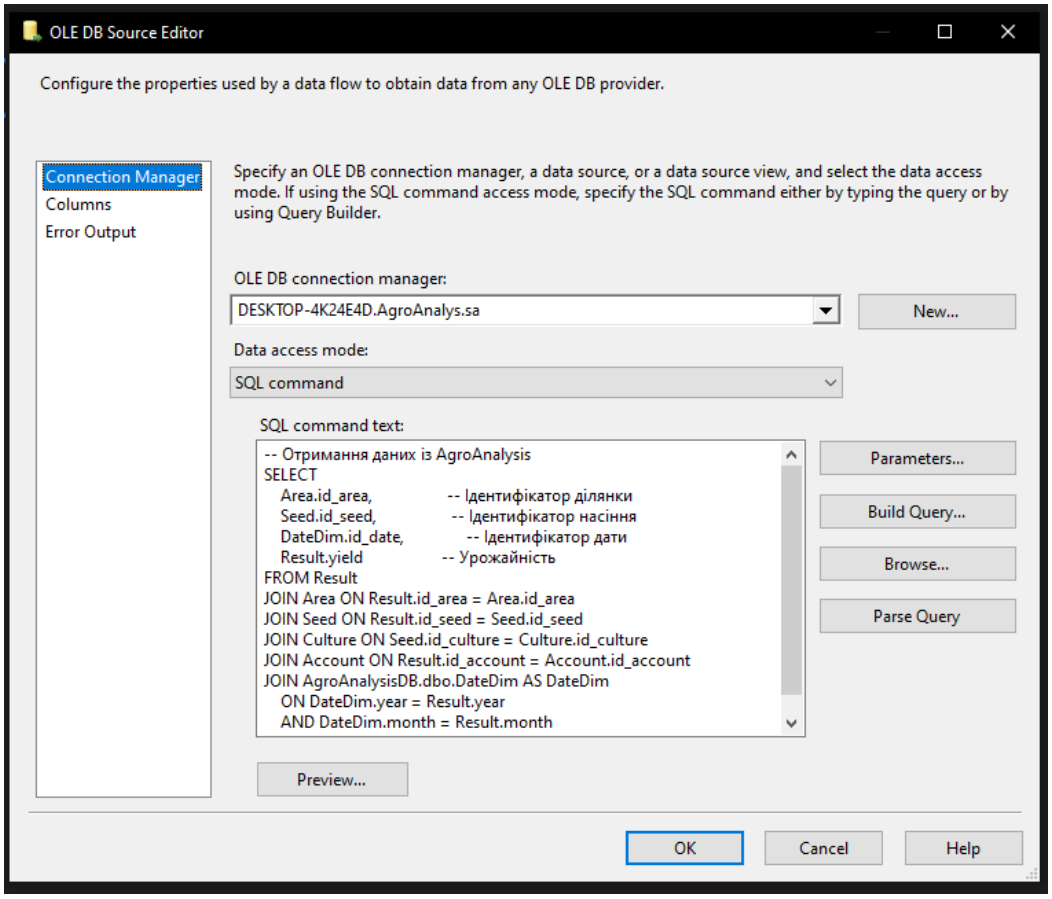


Рисунок 3.6.6 Об’єднувальний SQL-запит для факту

На рис. 3.6.7 відображено успішне завершення етапу Data Flow: для джерела OLE DB Result і приймача OLE DB Fact показано статус Success та підсумковий лічильник переданих рядків (у наведеному прикладі 3 rows). Це означає, що код пройшов усі перевірки мапінгу колонок, типів даних і зовнішніх ключів, а також не згенерував помилок чи відхилених записів у гілці Error Output. Додатково у вікні Execution Results фіксується час виконання кожного таска, що дозволяє контролювати продуктивність і порівнювати тривалість виконання при оновленнях.

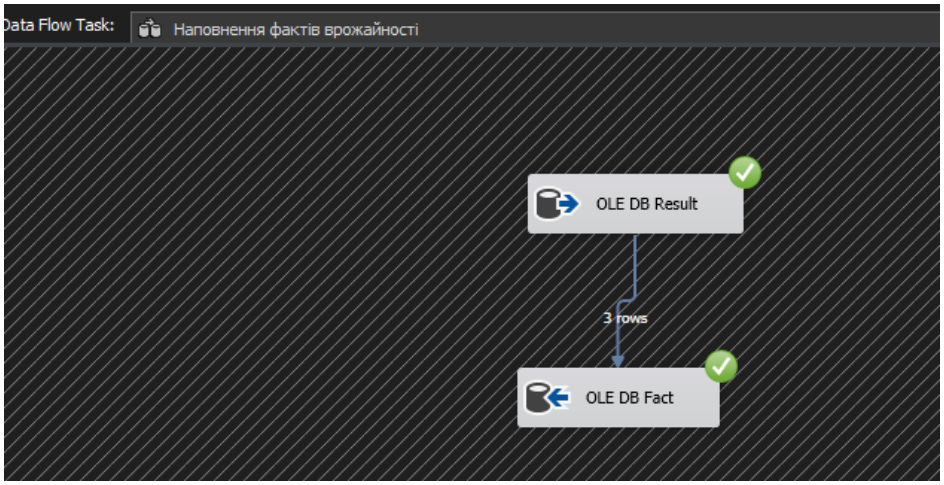


Рисунок 3.6.7 Результат виконання потоку фактів

На рис. 3.6.8 показано заповнений вимір Dim_Area: для кожної області сформовані записи з унікальним id_area та атрибутами area_name, soil_type, climate (наприклад, Вінницька - чорнозем типовий - помірно-континентальний). Видно, що ключі не дублюються, відсутні NULL у критичних полях і коректно збережено українські назви. Аналогічні контрольні вибірки виконано для інших вимірів і таблиці фактів.

	id_area	area_name	soil_type	climate
1	1	Вінницька	Чорнозем типовий	Помірно-континентальний
2	2	Волинська	Дерново-підзолистий	Вологий помірний
3	3	Дніпропетровська	Чорнозем звичайний	Континентальний
4	4	Донецька	Чорнозем південний	Посушливий
5	5	Житомирська	Поліський піщаний	Помірний вологий
6	6	Закарпатська	Бурозем	М'який вологий
7	7	Запорізька	Чорнозем південний	Континентальний сухий
8	8	Івано-Франківська	Бурозем	Вологий гірський
9	9	Київська	Сірий лісовий	Помірно-континентальний
10	10	Кіровоградська	Чорнозем типовий	Континентальний
11	11	Луганська	Каштановий	Сухий степовий
12	12	Львівська	Дерново-підзолистий	Помірний вологий
13	13	Миколаївська	Чорнозем південний	Сухий
14	14	Одеська	Чорнозем південний	Континентальний з посухами
15	15	Полтавська	Чорнозем типовий	Помірно-континентальний
16	16	Рівненська	Дерново-підзолистий	Помірний
17	17	Сумська	Сірий лісовий	Помірний вологий
18	18	Тернопільська	Чорнозем	Помірно-вологий
19	19	Харківська	Чорнозем звичайний	Помірно-континентальний
20	20	Херсонська	Каштановий	Посушливий
21	21	Хмельницька	Сірий лісовий	Вологий
22	22	Черкаська	Чорнозем типовий	Помірно-континентальний
23	23	Чернівецька	Бурозем	М'який вологий

Рисунок 3.6.8 Перевірка заповнення вимірів у СУБД

4 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ

4.1 Дослідження результатів застосування методу 1-Rule для класифікації даних

Вибірка сформована зі сховища AgroAnalysisDB. Із OLAP-куба отримано зрізи з такими полями: область (area), тип грунту (soil_type), кліматична ознака (climate), культура (culture), рік, урожайність (yield).

Для класифікації побудовано ціль YieldClass - H, L (High/Low) на основі середніх рівнів врожайності по культурах, які показані в таблиці 4.1.

Таблиця 4.1

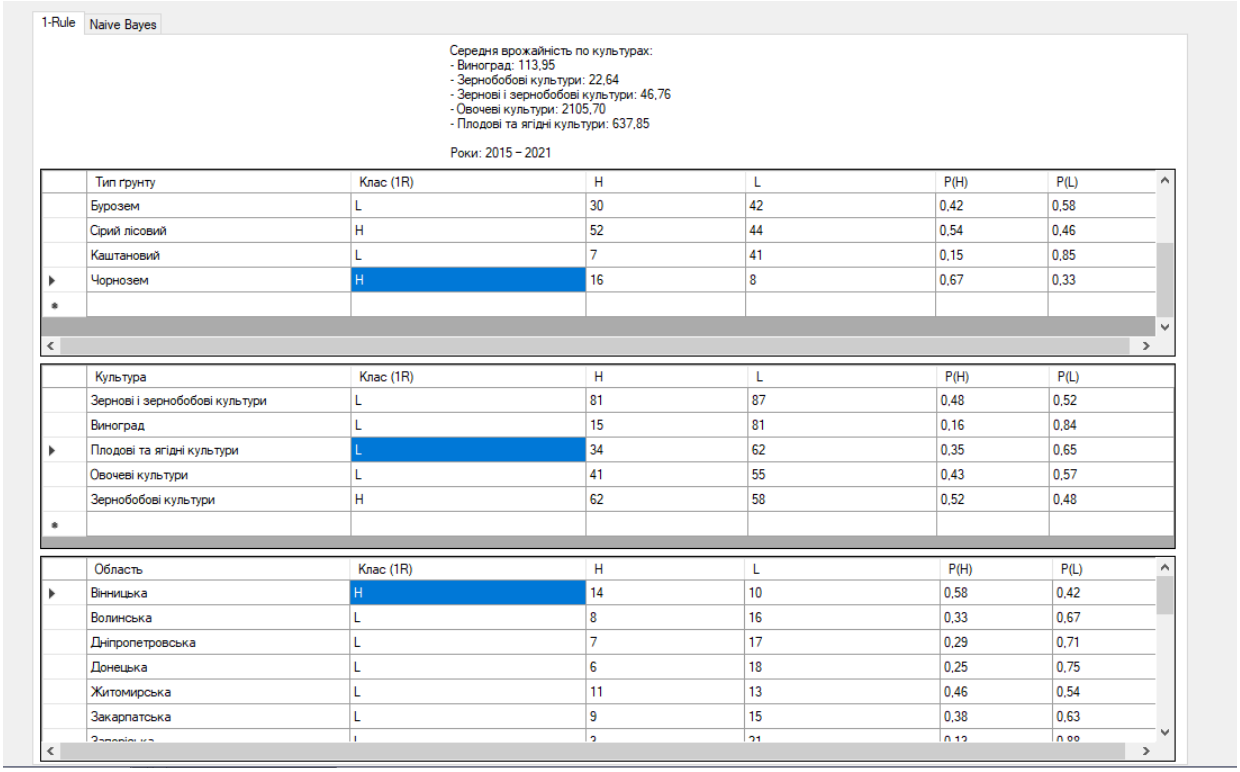
Середній рівень врожайності по культурам

Культура	Середній рівень врожайності
Виноград	113,95
Зернобобові	22,64
Зернові і зернобобові	46,76
Овочеві	2015,70
Плодові та ягідні	637,85

Спостереження з врожайністю вище порога для своєї культури позначалися як H, інакше - L.

Було оцінено три найпрактичніші ознаки окремо: тип грунту, культура, область. Для кожної з них 1R побудував таблицю рішень “значення - клас” та оцінив частоти/ймовірності класів.

На рис 4.1 подано обраний клас для кожного значення ознаки (стовпець Клас (1R)) та оцінені ймовірності P(H) і P(L).



Тип грунту	Клас (1R)	H	L	P(H)	P(L)
Бурозем	L	30	42	0,42	0,58
Сірий лісовий	H	52	44	0,54	0,46
Каштановий	L	7	41	0,15	0,85
Чорнозем	H	16	8	0,67	0,33

Культура	Клас (1R)	H	L	P(H)	P(L)
Зернові і зернобобові культури	L	81	87	0,48	0,52
Виноград	L	15	81	0,16	0,84
Плодові та ягідні культури	L	34	62	0,35	0,65
Овочеві культури	L	41	55	0,43	0,57
Зернобобові культури	H	62	58	0,52	0,48

Область	Клас (1R)	H	L	P(H)	P(L)
Вінницька	H	14	10	0,58	0,42
Волинська	L	8	16	0,33	0,67
Дніпропетровська	L	7	17	0,29	0,71
Донецька	L	6	18	0,25	0,75
Житомирська	L	11	13	0,46	0,54
Закарпатська	L	9	15	0,38	0,63

Рис 4.1 1-Rule: підсумкова панель за ознаками (тип грунту, культура, область).

- **Тип грунту.** Для чорноземів переважає високий клас врожайності (P(H) ≈ 0,67), тоді як для каштанових грунтів імовірність H значно нижча.
- **Культура.** Для овочевих культур 1R фіксує підвищений шанс високої врожайності (P(H) ≈ 0,43) відносно порогових значень своєї групи, зернобобові частіше відносяться до L.
- **Область.** За Вінницькою областю спостерігається вищий шанс класу H (P(H) ≈ 0,58), що узгоджується з агрокліматичними передумовами регіону.

1R чітко підсвічує вплив грунтових умов і географії на досягнення високих показників. Для первинного планування це дає прості правила:

- “якщо грунт = чорнозем, очікуємо рівень врожайності високий”;

- “якщо культура = овочеві, очікуємо рівнь врожайності вищий за середній”;
- “якщо область = Вінницька, то очікуємо рівень врожайності вищий за середній ”.

1R ігнорує взаємодії ознак (наприклад чорнозем * конкретна культура * рік). На точність впливають коливання погоди, агротехнічні практики й сезонні ефекти, які одна ознака не відображає. Тому 1R використано як основа та інструмент інтерпретації, а для багатофакторних залежностей у наступних підрозділах застосовано Наївний Байєс, Apriori та кластеризацію.

Алгоритм 1-Rule підтвердив домінуючу роль типу грунту та регіону для досягнення високої врожайності й окреслив точки уваги для подальшого аналізу: чорноземи, овочеві культури й група областей-лідерів. Отримані правила стали базою для швидких пояснюваних рекомендацій і валідації складніших моделей.

4.2 Використання алгоритму Наївний Байєс

Навчальна вибірка сформована зі сховища AgroAnalysisDB. Використано категоріальні ознаки:

- Область (географія вирощування)
- Культура
- Тип грунту

Цільова змінна YieldClass - {H, L} отримана співставленням урожайності відносно середніх порогів по культурах (див. рис. 4.2). Ознаки кодувалися one-hot, значення перевірялися на повноту й логічну узгодженість (відсутність від’ємних величин, коректні коди регіонів і культур).

1-Rule | Naive Bayes

Область	Культура	Тип грунту	Клас фактичний	Клас (Naive Bayes)	P(H)	P(L)
Вінницька	Зернові і зернобо...	Чорнозем типо...	H	H	0,71	0,29
Вінницька	Зернові і зернобо...	Чорнозем типо...	H	H	0,71	0,29
Вінницька	Виноград	Чорнозем типо...	L	L	0,34	0,66
Вінницька	Плодові та ягідн...	Чорнозем типо...	H	H	0,59	0,41
Вінницька	Овочеві культури	Чорнозем типо...	L	H	0,66	0,34
Вінницька	Зернобобові кул...	Чорнозем типо...	H	H	0,73	0,27
Вінницька	Зернові і зернобо...	Чорнозем типо...	H	H	0,71	0,29
Вінницька	Зернобобові кул...	Чорнозем типо...	L	H	0,73	0,27
Вінницька	Зернові і зернобо...	Чорнозем типо...	H	H	0,71	0,29
Вінницька	Виноград	Чорнозем типо...	L	L	0,34	0,66
Вінницька	Плодові та ягідн...	Чорнозем типо...	L	H	0,59	0,41
Вінницька	Овочеві культури	Чорнозем типо...	L	H	0,66	0,34
Вінницька	Зернобобові кул...	Чорнозем типо...	L	H	0,73	0,27
Вінницька	Зернові і зернобо...	Чорнозем типо...	H	H	0,71	0,29
Вінницька	Виноград	Чорнозем типо...	L	L	0,34	0,66
Вінницька	Плодові та ягідн...	Чорнозем типо...	H	H	0,59	0,41
Вінницька	Овочеві культури	Чорнозем типо...	H	H	0,66	0,34
Вінницька	Зернобобові кул...	Чорнозем типо...	L	H	0,73	0,27
Вінницька	Зернові і зернобо...	Чорнозем типо...	H	H	0,71	0,29
Вінницька	Виноград	Чорнозем типо...	L	L	0,34	0,66
Вінницька	Плодові та ягідн...	Чорнозем типо...	H	H	0,59	0,41
Вінницька	Овочеві культури	Чорнозем типо...	H	H	0,66	0,34
Вінницька	Зернобобові кул...	Чорнозем типо...	H	H	0,73	0,27
Вінницька	Зернові і зернобо...	Чорнозем типо...	H	H	0,71	0,29
Волинська	Зернові і зернобо...	Дерново-підзол...	L	L	0,46	0,54
Волинська	Зернові і зернобо...	Дерново-підзол...	L	L	0,46	0,54
Волинська	Виноград	Дерново-підзол...	L	L	0,15	0,85

Рис 4.2 Наївний Байєс: імовірності

На рис. 4.2 наведено зведену таблицю з колонками: Область, Культура, Тип грунту, Клас фактичний, Клас (Naive Bayes), P(H), P(L). З прикладів на знімку:

- Вінницька область * Зернові і зернобобові * Чорнозем типовий: P(H)=0,71,P(L)=0,29 - високий шанс досягти класу H
- Вінницька область * Виноград * Чорнозем типовий: P(H)=0,34 - імовірність H значно нижча порівняно з зерновими
- Вінницька область * Овочеві культури * Чорнозем типови: P(H)≈0,66 – стабільно високі шанси

Для кожної комбінації обчислено ймовірності високої та низької врожайності (P(H) і P(L)) і наведено порівняння фактичного класу з передбаченим моделлю. Це дозволяє оцінити не лише сам прогноз, а й ступінь упевненості моделі для конкретних агроумов.

Метод Наївного Байєса показує, що результат визначається поєднанням факторів, а не однією ознакою. Для чорноземів у зернових у Вінницькій

області шанс H суттєво вищий, ніж для винограду на тих самих грунтах - отже, культура * грунт * регіон взаємодіють.

Висока ймовірність P(H) означає, що за вказаних умов культура має високі шанси дати врожайність вище порогового рівня. Це можна застосувати для:

- відбору культур у регіонах: обирати комбінації регіон-культура-грунт з найвищою P(H) як основні сценарії посівів;
- виявлення ризикових зон: комбінації з низькою P(H) вважати підвищеним ризиком і коригувати площі або змінювати культуру/технологію;
- коригування ресурсів: у зонах із проміжними значеннями P(H) планувати додаткове зрошення, добрива чи інші заходи, щоб підвищити шанси на клас H.

Припущення про незалежність ознак порушується частково (географія корелює з грунтами/кліматом). Це може занижувати/завищувати оцінки для деяких комбінацій. Для контролю якості порівнюємо передбачений клас із фактичним у тому ж зрізі (колонки Клас фактичний та Клас (Naive Bayes) а також відстежуємо частку правильних передбачень у часі.

Наївний Байєс надав кількісні, інтерпретовані ймовірності високої врожайності для конкретних комбінацій область-культура-тип грунту та перевершив 1-Rule у здатності враховувати декілька факторів одночасно. Ці оцінки використано як шар прогноз/впевненість, що доповнює базові KPI і допомагає приймати рішення щодо розміщення культур та планування технологій.

4.3 Застосування алгоритму Apriori

Із OLAP-куба було сформовано бінаризовані ознаки (one-hot): регіон/область, тип грунту, кліматична характеристика, культура та цільова подія Врожайність=Висока (H). Пороги для H/L - ті самі, що застосовувались

у 4.1–4.2 (за середніми по культурах). Для відсікання шуму виставлено мінімальні пороги support і confidence. Далі виконано прорідження дублетів (правил, що повторюють зміст з надлишковими умовами).

Після обробки транзакцій алгоритм згенерував >20 правил. Значення метрик:

- support - ~2–9% (частка записів, де правило спрацьовує)
- confidence - 0.5–1.0 (ймовірність Н за виконання умов)
- lift – >1 (позитивна залежність), у низці випадків – дуже високий (до ≈13), що вказує на сильні локальні закономірності

На рис. 4.3. наведено Вибірка згенерованих правил Apriori з метриками підтримка, ймовірність (довіра) та ліфт. На фрагменті видно як однофакторні правила (клімат/регіон/культура), так і комбінації умов.

Умова	Результат	Підтримка	Ймовірність	Ліфт (асоціація)
frozenset({'Клімат=Вологий'})	frozenset({'Врожайність=Висока'})	0,027777778	0,666666667	1,324137931
frozenset({'Клімат=Континентальний з посуками'})	frozenset({'Врожайність=Висока'})	0,020833333	0,5	0,993103448
frozenset({'Клімат=Помірний вологий'})	frozenset({'Врожайність=Висока'})	0,088541667	0,53125	1,055172414
frozenset({'Клімат=Помірно-вологий'})	frozenset({'Врожайність=Висока'})	0,026041667	0,625	1,24137931
frozenset({'Клімат=Помірно-континентальний'})	frozenset({'Врожайність=Висока'})	0,114583333	0,55	1,092413793
frozenset({'Клімат=Сухий'})	frozenset({'Врожайність=Висока'})	0,020833333	0,5	0,993103448
frozenset({'Культура=Овочеві культури'})	frozenset({'Врожайність=Висока'})	0,166666667	1	1,986206897
frozenset({'Культура=Плодові та ягідні культури'})	frozenset({'Врожайність=Висока'})	0,166666667	1	1,986206897
frozenset({'Регіон=Вінницька'})	frozenset({'Врожайність=Висока'})	0,024305556	0,583333333	1,15862069
frozenset({'Регіон=Дніпропетровська'})	frozenset({'Врожайність=Висока'})	0,020833333	0,5	0,993103448
frozenset({'Регіон=Закарпатська'})	frozenset({'Врожайність=Висока'})	0,020833333	0,5	0,993103448
frozenset({'Регіон=Київська'})	frozenset({'Врожайність=Висока'})	0,024305556	0,583333333	1,15862069
frozenset({'Регіон=Миколаївська'})	frozenset({'Врожайність=Висока'})	0,020833333	0,5	0,993103448
frozenset({'Регіон=Одеська'})	frozenset({'Врожайність=Висока'})	0,020833333	0,5	0,993103448
frozenset({'Регіон=Полтавська'})	frozenset({'Врожайність=Висока'})	0,026041667	0,625	1,24137931
frozenset({'Регіон=Сумська'})	frozenset({'Врожайність=Висока'})	0,026041667	0,625	1,24137931
frozenset({'Регіон=Тернопільська'})	frozenset({'Врожайність=Висока'})	0,026041667	0,625	1,24137931
frozenset({'Регіон=Херсонська'})	frozenset({'Врожайність=Висока'})	0,020833333	0,5	0,993103448
frozenset({'Регіон=Хмельницька'})	frozenset({'Врожайність=Висока'})	0,027777778	0,666666667	1,324137931
frozenset({'Регіон=Черкаська'})	frozenset({'Врожайність=Висока'})	0,024305556	0,583333333	1,15862069
frozenset({'Регіон=Чернігівська'})	frozenset({'Врожайність=Висока'})	0,024305556	0,583333333	1,15862069
frozenset({'Грунт=Сірий лісовий'})	frozenset({'Врожайність=Висока'})	0,102430556	0,614583333	1,220689655
frozenset({'Грунт=Чорнозем'})	frozenset({'Врожайність=Висока'})	0,026041667	0,625	1,24137931
frozenset({'Грунт=Чорнозем типовий'})	frozenset({'Врожайність=Висока'})	0,092013889	0,552083333	1,096551724
frozenset({'Культура=Зернобобові культури'})	frozenset({'Врожайність=Низька'})	0,140625	0,675	3,135483871
frozenset({'Культура=Зернові і зернобобові культури'})	frozenset({'Врожайність=Середня'})	0,189236111	0,648809524	2,306878307
frozenset({'Клімат=Вологий', 'Культура=Зернові і зернобобові культури'})	frozenset({'Врожайність=Висока'})	0,012152778	1	1,986206897
frozenset({'Клімат=Вологий', 'Регіон=Хмельницька'})	frozenset({'Врожайність=Висока'})	0,027777778	0,666666667	1,324137931
frozenset({'Клімат=Вологий'})	frozenset({'Врожайність=Висока', 'Регіон=Хмел	0,027777778	0,666666667	24
frozenset({'Регіон=Хмельницька'})	frozenset({'Клімат=Вологий', 'Врожайність=Вис	0,027777778	0,666666667	24
frozenset({'Клімат=Вологий', 'Грунт=Сірий лісовий'})	frozenset({'Врожайність=Висока'})	0,027777778	0,666666667	1,324137931
frozenset({'Клімат=Вологий'})	frozenset({'Грунт=Сірий лісовий', 'Врожайність	0,027777778	0,666666667	6,508474576
frozenset({'Клімат=Континентальний', 'Культура=Овочеві культури'})	frozenset({'Врожайність=Висока'})	0,013888889	1	1,986206897
frozenset({'Культура=Плодові та ягідні культури', 'Клімат=Континентальний'})	frozenset({'Врожайність=Висока'})	0,013888889	1	1,986206897
frozenset({'Регіон=Дніпропетровська', 'Клімат=Континентальний'})	frozenset({'Врожайність=Висока'})	0,020833333	0,5	0,993103448
frozenset({'Регіон=Дніпропетровська'})	frozenset({'Клімат=Континентальний', 'Врожай	0,020833333	0,5	13,09090909

Рис 4.3 Приклади інтерпретацій

- Регіон=Дніпропетровська і Клімат=Континентальний - Врожайність=Висока - lift ~13: комбінація істотно підвищує шанс класу Н порівняно із загальним рівнем.

- Клімат=Вологий - Врожайність=Висока - confidence близько 0.67: вологий клімат у середньому підштовхує показник у зону високих значень.
- Грунт=Чорнозем типовий - Врожайність=Висока - стабільний позитивний зв’язок, ліфт >1 підтверджує перевагу чорноземів.
- Комбіновані умови Регіон + Грунт + Культура демонструють, що поєднання факторів часто важливіше за будь-який із них окремо.

Практичний сенс правил.

- Вибір культур: у регіонах і на грунтах із правилами з високим lift доцільно надати пріоритет відповідній культурі (очікуваний шанс H вищий).
- Карта ризиків: правила на L (якщо присутні) виділяють небажані сценарії - корисно для планування зрошення, добрив, строків.
- Персоналізація технологій: комбінації клімат * грунт * культура підказують, де інтенсивні технології дадуть найбільший ефект.

Apriori доповнив 1R і Наївний Байєс людинозрозумілими якщо-то правилами, показавши, у яких поєднаннях регіон–клімат–грунт–культура ймовірність високої врожайності суттєво вища за середню.

4.4 Кластерний аналіз

4.4.1 Вибір кількості кластерів (Elbow Method)

Для алгоритму K-Means кількість кластерів K задається наперед: замалий K зливає різні типи об’єктів, завеликий - дробить дані та погіршує узагальнюваність. Тому K підбирали емпірично за методом ліктя.

Для K=1…10 навчали K-Means (ініціалізація k-means++, зафіксований random_state), для кожного значення обчислювали inertia (внутрішньокластерну дисперсію). Далі будували графік і шукали точку, після якої додаткові кластери зменшують помилку незначно.

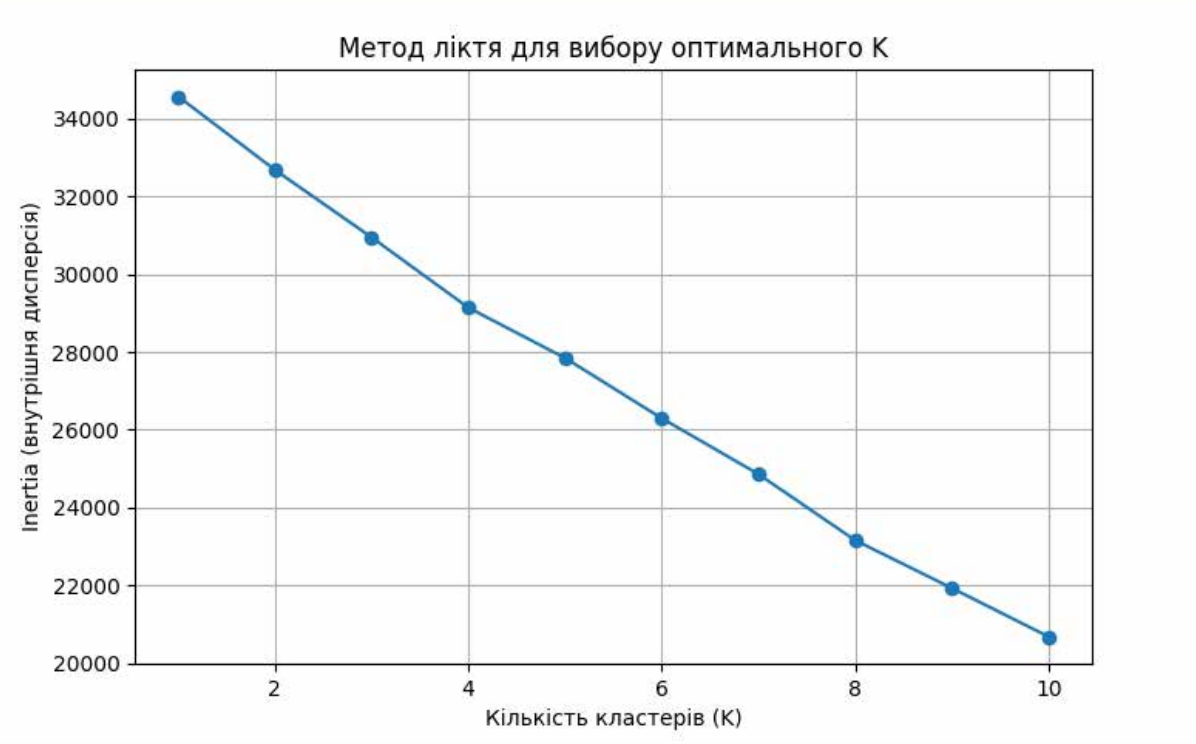


Рис 4.4.1 Метод ліктя для вибору оптимального K.

На ділянці K=1–4 дисперсія різко падає. Після K≈4 крива згладжується – формується лікоть. Оптимальним прийнято K=4 як баланс між деталізацією та узагальненням. Саме з цим значенням виконано фінальну кластеризацію датасету.

4.4.2 Візуалізація та інтерпретація кластерів (PCA)

Щоб перевірити якість розділення та зрозуміти структуру груп, результати K-Means (за K=4) спроєктовано у 2-вимірний простір методом головних компонент (PCA).

“PCA - це підхід для зменшення розмірності великих наборів даних шляхом перетворення великого набору змінних на менший, який зберігає більшу частину інформації з великого набору.”[7]

Перші дві компоненти акумулюють найбільшу частку дисперсії, тож взаємне розташування кластерів інформативне.

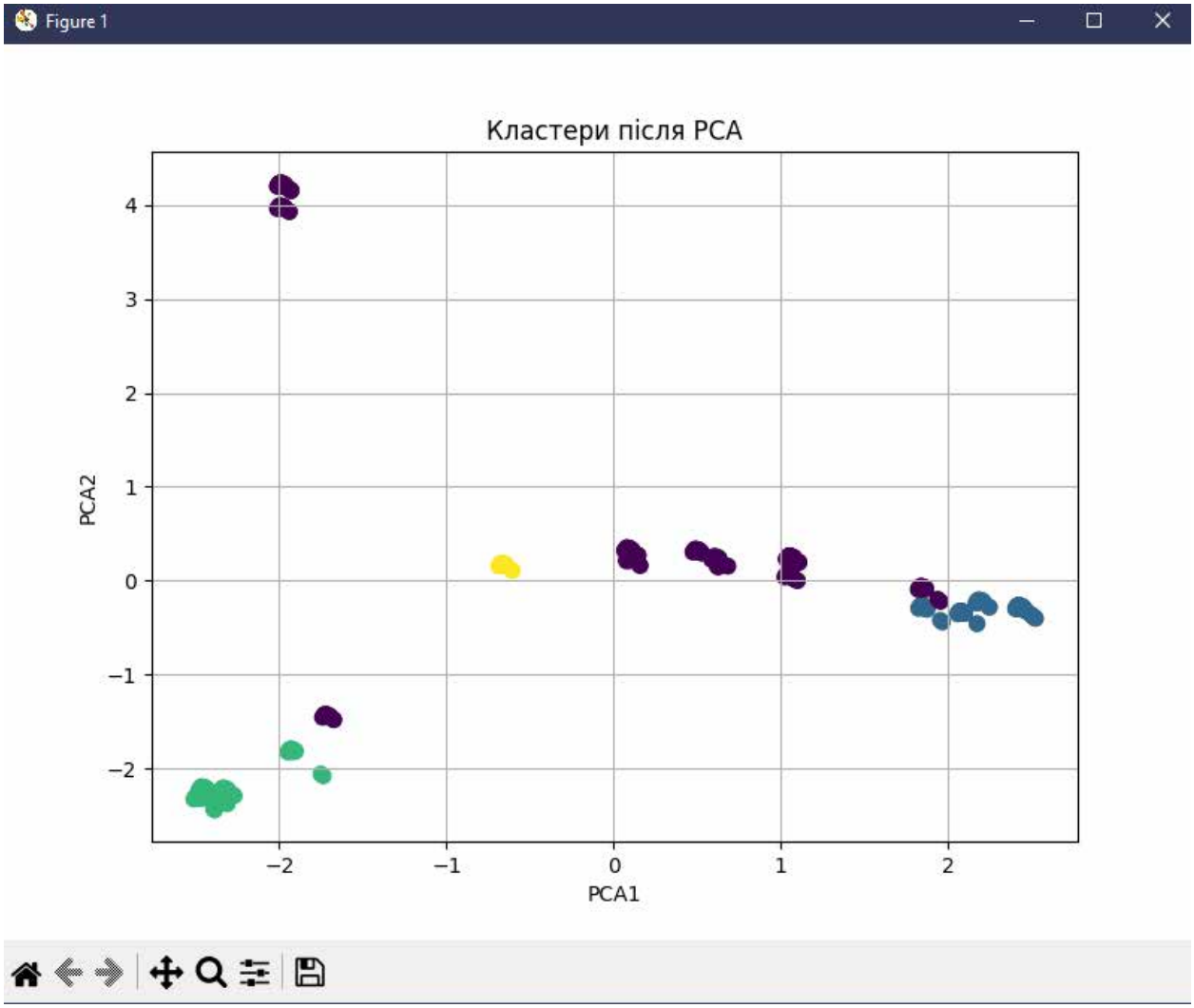


Рис 4.4.2 Кластери після PCA

Рис. 4.4.2 кожна точка - запис, колір - мітка кластера. Спостерігаються компактні, здебільшого чітко відокремлені хмари, мінімальне перекриття лише на межах.

Компактність хмар підтверджує низьку внутрішньокластерну дисперсію - ціль K-Means досягнуто. Рознесеність центрів у площині PCA свідчить, що кластери відрізняються комбінаціями ознак (тип грунту, клімат, культура, профіль врожайності).

Узагальнені профілі кластерів.

- Кластер A (високий і стабільний потенціал): переважають чорноземи та помірно-вологі умови, зернові/овочеві, середня врожайність вище загальної, стабільність висока.
- Кластер B (нестабільний високий): більша частка плодових/ягідних, чутливість до року, потребує керування ризиками (зрошення, строки).
- Кластер C (нижчий потенціал): континентальніші умови, частка зернобобових, середня врожайність нижча, доцільні адресні технологічні втручання.
- Кластер D (середній, але стабільний): змішані грунти, часто континентальний клімат, показники близькі до середніх, зате коливання невеликі.

Поєднання Elbow Method (для коректного вибору K) і PCA (для інтерпретації) підтвердило наявність природної сегментації агроданих на чотири стійкі групи. Це забезпечує надійну основу для подальшої аналітики, формування рекомендацій та прийняття управлінських рішень у рослинництві.

4.5 КПІ

KPI узгоджують результати OLAP-аналізу та Data Mining із управлінськими цілями. Вони показують, де ми просідаємо системно (середній рівень по країні/областях), а де вже є еталонні практики (топ-культури, топ-грунти) для масштабування.

Усі значення KPI отримуються безпосередньо з OLAP-куба (міра yield, ієрархії Рік, Область, Культура, Тип грунту). Агрегування - середнє по відповідних групах, для Top застосовується пошук максимуму серед середніх значень (TOP-N). Кожен КПІ описаний в таблиці 4.5.

Таблиця 4.5

Опис КПІ

KPI	Базове значення	Цільове значення	Пояснення вибору цілі
Avg Yield by Area	середня врожайність по областях (494,61 ц/га)	2*базове (989,22 ц/га)	Підвищити середній рівень до двох середніх - амбіційна, але досяжна мета для загального росту.
Top Crop Yield	максимальна середня врожайність культури (2105,70 ц/га)	1,5*базове (3158,55 ц/га)	Ставимо помірнішу, ніж у попередньому KPI, мету - удосконалити вже найкращий результат на 50 %.
Top Soil Yield	максимальна середня врожайність по типам грунту (709,40 ц/га)	1,2*базове (851,28 ц/га)	Поліпшити продуктивність найкращих грунтів на 20 %.

Зони інтерпретації:

- Червона: 0–50 % від цілі - критично, потрібні негайні дії.
- Жовта: 50–75 % - прийнятно, але недостатньо.
- Зелена: 75–100 % - близько до мети або перевиконання.

Візуалізація KPI

Рис 4.5.1 Гейджі Avg Yield by Area та Top Crop Yield

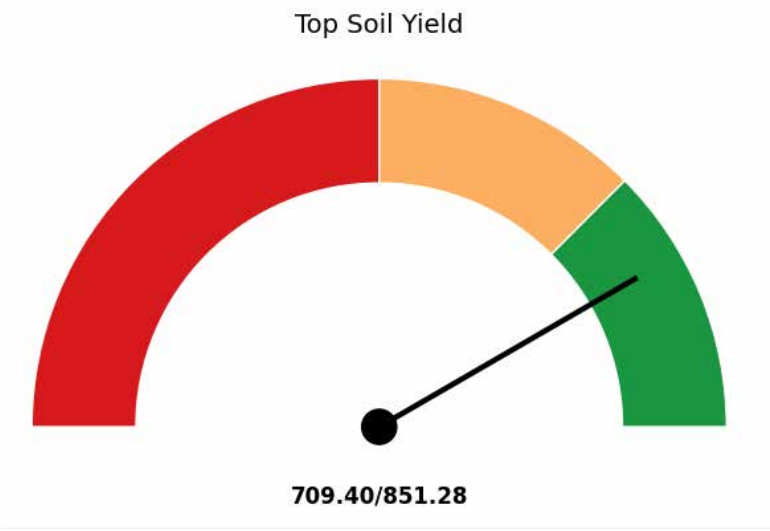


Рис 4.5.2 Гейдж Top Soil Yield

Тлумачення поточного стану

- Avg Yield by Area

  Вимірює: середню врожайність усіх культур у розрізі областей.

  Факт/Ціль: 494,61 / 989,22 ц/га ≈ 50 % до мети.

Висновок: стрілка на межі червоного/жовтого - потрібні системні покращення (технології, строки сівби, живлення, зрошення) у широкому колі регіонів.

- Top Crop Yield

  Вимірює: середню врожайність найпродуктивнішої культури.

  Факт/Ціль: 2 105,70 / 3 158,55 ц/га ≈ 67 % до мети.

  Висновок: жовта зона - є резерв росту через масштабування практик із регіонів-лідерів (див. правила Apriori та кластер високий і стабільний).

- Top Soil Yield

  Вимірює: середню врожайність на найкращих типах грунтів.

  Факт/Ціль: 709,40 / 851,28 ц/га ≈ 83 % до мети.

  Висновок: зелена зона - висока ефективність локацій із топ-грунтами. Доцільно розширювати площі культур-лідерів у цих зонах.

4.6 Зв’язок KPI із Data Mining

У межах проекту ключові показники ефективності відображають поточний стан системи вирощування культур у різних розрізах (час, регіон, культура, грунт). Водночас методи інтелектуального аналізу даних забезпечують причинно-пояснювальну складову. Вони виявляють закономірності, оцінюють імовірності досягнення цільових значень і пропонують керовані важелі впливу.

У системі застосовано три базові KPI. Показник “Avg Yield by Area” відображає середню врожайність у регіональному розрізі та використовується для оцінки загального рівня результативності. “Top Crop Yield” фіксує середню врожайність найпродуктивнішої культури та дозволяє відстежувати прогрес лідера і можливості його масштабування. “Top Soil Yield” характеризує середню врожайність на найкращих за продуктивністю типах грунту й служить орієнтиром ефективності використання грунтового потенціалу. Усі три показники агрегуються у кубі OLAP за узгодженими

формулами та доповнюються цільовими рівнями й порогами інтерпретації (червона/жовта/зелена зони).

Результати Data Mining підкріплюють KPI конкретними рекомендаціями. Асоціативний аналіз (Apriori) формує правила, для яких обчислюються підтримка, довіра та ліфт. Саме такі поєднання можуть бути використані як важелі підвищення “Avg Yield by Area”: на регіональних зрізах відбираються ділянки з сумісним профілем грунту та клімату і планується розширення посівів відповідних культур.

Модель Naive Bayes оцінює ймовірності класів і дає ймовірність ризику та потенціалу, що доповнюють загальну картину. Для кожної комбінації область-культура-тип ґрунту модель обчислює P(H) і P(L) та пропонує прогнозований клас. Високі значення P(H) свідчать про перспективність розширення площ під конкретні культури в конкретних регіонах,що безпосередньо підтримує зростання КПІ “Top Crop Yield”, тоді як поєднання з низькою P(H) маркуються як ризикові та потребують зміни культури або посилення технологій.

Кластеризація виокремлює кластери з близькими характеристиками. Кластери з високою та стабільною врожайністю позначаються як зони пріоритетного масштабування: для них планується розширення площ і копіювання технологій. Групи з високою, але нестабільною врожайністю розглядаються як зони керування ризиком: запроваджуються можливі покращення, аби не втратити потенціал. Кластери зі середніми чи низькими значеннями використовуються для технологічних змін.

KPI надають цілісне уявлення про стан системи, а Data Mining дає конкретні причини та точки прикладання зусиль. Завдяки цьому управлінські дії спираються не на середні оцінки по системі, а на локальні закономірності, підтверджені підтримкою, довірою, ліфтом, імовірностями та стабільністю кластерів, що у підсумку забезпечує більш прогнозоване зростання показників урожайності.

ВИСНОВКИ

У цій магістерській роботі продемонстровано, що поєднання сховища даних та OLAP із методами інтелектуального аналізу формує цілісний інструмент для оцінювання й прогнозування врожайності. Створена інфраструктура - схема “зірка” з фактом врожайності та вимірами час–регіон–грунт–культура–насіння–посів, SSIS-конвеєр завантаження та куб SSAS - забезпечує стабільний і відтворюваний цикл: від збирання й узгодження джерел до інтерактивних зрізів, моделей і управлінських показників.

OLAP-рівень дав оперативний огляд ситуації: зрізи за роками, областями, культурами та типами грунтів. Картина, що склалася на гейджах, є контрастною: середня врожайність по регіонах відстає від цільового коридору, водночас показники топ-культур і топ-грунтів наближаються до планових меж. Це вказує на потребу піднімати середній рівень через масштабування практик із локальних точок успіху.

Data Mining перетворив описові спостереження на формалізовані висновки:

- 1R висвітив найсильніші однофакторні сигнали (зокрема перевагу чорноземів для високих класів врожайності) - як орієнтири для подальшого аналізу
- Наївний Байєс надав імовірнісні оцінки для комбінацій регіон–культура–тип грунту, для частини зернових зафіксовано P(H) на рівні ≈0,7, тоді як для культур, менш сумісних із локальними умовами, імовірність високої врожайності помітно нижча. Таким чином, вирішальним виявився профіль поєднання факторів, а не окрема ознака
- Apriori виявив правила з адекватною підтримкою та високим ліфтом (до ≈13), що чітко позначають

сприятливі зв'язки клімат–регіон–культура і можуть безпосередньо використовуватися як рекомендації для розміщення культур

- K-Means із добором K методом ліктя (оптимум K=4) та подальшою PCA-проєкцією показав природну сегментацію даних на кластери з відмінними агропрофілями: від високих і стабільних до середніх, але стійких. Це дає основу для диференційованих агротехнологій і пріоритетів інвестицій

OLAP швидко локалізує проблемні та перспективні зони, моделі вказують, чому саме там спостерігається спад чи зростання, KPI фіксують прогрес щодо цілей. Разом це переводить роботу з даними у цикл: виявити - пояснити - обрати дію - перевірити.

Залежність від повноти та якості відкритих даних, часткове порушення припущення незалежності ознак у Naive Bayes, локальність частини асоціацій із високим ліфтом і невеликою підтримкою. Вони задають напрям подальшої роботи: розширення джерел (польові сенсори, супутникові індекси, агрохімічні карти), посилення прогнозування часовими моделями й каузальним аналізом, а також глибша інтеграція рекомендацій у операційні процеси.

Розроблена система довела спроможність перетворювати різнорідні агродані на рішення, що мають практичний сенс: від вибору культур під конкретні грунтово-кліматичні умови до планування ресурсів і контролю виконання цілей. Основну мету дослідження досягнуто, отриманий інструментарій придатний до практичної перевірки та подальшого розвитку.

ДЖЕРЕЛА

1. Державна служба статистики України. Форма № 29-сг (річна): “Звіт про площі та валові збори сільськогосподарських культур…”. URL: https://www.ukrstat.gov.ua/norm_doc/2024/91/29-sg_rik.pdf (дата звернення: 20.06.2025).
2. Державна служба статистики України. Роз’яснення щодо заповнення форми № 29-сг (річна). URL: https://ukrstat.gov.ua/albom/albom_2020/roz_2020/2.03.07/roz_29_sg_17.doc (дата звернення: 20.06.2025).
3. Державна служба статистики України. Роз’яснення щодо форми № 37-сг (місячна). URL: https://www.ukrstat.gov.ua/albom/albom_2025/2.03.07/roz_37_sg_08.07.2022.doc (дата звернення: 26.06.2025).
4. Міністерство цифрової трансформації України. “Кодифікатор адміністративно-територіальних одиниць та територій територіальних громад (КАТОТТГ)” (відкриті дані). URL: https://data.gov.ua/dataset/43c2a113-a032-4c8a-a409-3b5e1660bb38 (дата звернення: 26.06.2025).
5. Що таке алгоритм Апріорі в аналізі даних?. URL: https://www.maxzosim.com/sequence-diagrams (дата звернення: 15.07.2025).
6. Діаграма послідовності (Sequence Diagrams). URL: https://swim.liberty.cx.ua/psikhologiya/shho-take-algoritm-apriori-v-analizi-danikh.html (дата звернення: 15.07.2025).
7. PCA - Principal Component Analysis. URL: https://itwiki.dev/data-science/ml-reference/ml-glossary/pca-principal-component-analysis (дата звернення: 15.07.2025).

8. UML для бізнес-моделювання: для чого потрібні діаграми процесів. URL: https://evergreens.com.ua/ua/articles/uml-diagrams.html (дата звернення: 18.07.2025).
9. Microsoft Learn. “SQL Server Integration Services (SSIS)” - платформа для ETL. URL: https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver17 (дата звернення: 12.07.2025).
10. Microsoft Learn. “Key Performance Indicator (KPI) visuals - Power BI”. URL: https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-kpi (дата звернення: 13.08.2025).
11. Діаграма розгортання: Підручник з UML із ПРИКЛАДОМ. URL: https://www.guru99.com/uk/deployment-diagram-uml-example.html (дата звернення: 14.08.2025).
12. Що таке OLAP? Куб, аналітичний Operay сховищі даних. URL: https://www.guru99.com/uk/online-analytical-processing.html (дата звернення: 17.08.2025).
13. Що таке data mining?. URL: https://futurenow.com.ua/shho-take-data-mining-analiz-danyh (дата звернення: 22.08.2025).
14. Scikit-learn User Guide. “1.9. Naive Bayes”. URL: https://scikit-learn.org/stable/modules/naive_bayes.html (дата звернення: 18.09.2025).
15. Наївний алгоритм Байєса в машинному навчанні. URL: https://www.guru99.com/uk/naive-bayes-classifiers.html (дата звернення: 22.09.2025).
16. Орленко Н.С., Карпич М.К., Коховська І.В. Особливості сховища даних та оброблення результатів кваліфікаційної експертизи сортів рослин. URL: https://www.tnv-agro.ksauniv.ks.ua/archives/101_2018/15.pdf (дата звернення: 03.11.2025).

17. Climate FieldView. URL: https://www.climatefieldview.com.ua (дата звернення: 12.07.2025).
18. Основи комп'ютерного моделювання: поняття та методи. URL: https://optima.college/news/osnovi-komp-yuternogo-modelyuvannya-ponyattya-ta-metodi (дата звернення: 12.07.2025).
19. ETL-процес: написання вимог. URL: https://www.artofba.com/uk/post/etl-process-for-business-analyst (дата звернення: 18.08.2025).
20. Кращі практики моделювання даних за схема “зірка” схемою на Databricks SQL. URL: https://data-life-ua.com/modelling/krashchi-praktyky-modeliuvannia-danykh-za-star-schema-na-databricks-sql (дата звернення: 09.11.2025).
21. Видобуток даних (Data Mining). URL: https://www.maxzosim.com/data-mining (дата звернення: 15.05.2025).
22. Аналітична обробка в реальному часі. URL: https://vue.gov.ua/Аналітична_обробка_в_реальному_часі (дата звернення: 15.05.2025).
23. Що таке сховище даних? Типи, визначення та приклад. URL: https://www.guru99.com/uk/data-warehousing.html (дата звернення: 02.06.2025).
24. Що таке діаграма класів UML і найкращий творець діаграм класів UML. URL: https://www.mindonmap.com/uk/blog/what-is-uml-class-diagram (дата звернення: 07.07.2025).
25. Як будувати UML-діаграми. Розбираємо три найпопулярніші варіанти. URL: https://dou.ua/forums/topic/40575 (дата звернення: 11.08.2025).

ДОДАТКИ

ДОДАТОК А

код SQL:

```
-- Створення бази даних
CREATE DATABASE AgroAnalysisDB;
GO

-- Використання бази даних
USE AgroAnalysisDB;
GO

-- Таблиця вимірів для посівів
CREATE TABLE Dim_Sowing (
   id_sowing INT PRIMARY KEY,
   name_sowing NVARCHAR(100),
   description_sowing NVARCHAR(255),
);

-- Таблиця вимірів для ділянок
CREATE TABLE Dim_Area (
   id_area INT PRIMARY KEY,
   id_sowing INT,          -- Зв'язок із таблицею посівів
   soil_type NVARCHAR(50), -- Тип грунту
   climate NVARCHAR(50),   -- Кліматичні умови
   area_name NVARCHAR(100),-- Назва ділянки
   FOREIGN KEY (id_sowing) REFERENCES Dim_Sowing(id_sowing)
);

-- Таблиця вимірів для культур
CREATE TABLE Dim_Culture (
   id_culture INT PRIMARY KEY,
   name_culture NVARCHAR(100),
);
```

```
-- Таблиця вимірів для насіння
CREATE TABLE Dim_Seed (
    id_seed INT PRIMARY KEY,
    name_seed NVARCHAR(100),
    id_culture INT,
    productivity DECIMAL(10, 2), -- Продуктивність сорту
    FOREIGN KEY (id_culture) REFERENCES Dim_Culture(id_culture)
);

-- Таблиця вимірів для дати
CREATE TABLE DateDim (
    id_date INT IDENTITY(1,1) NOT NULL PRIMARY KEY,
    year INT NOT NULL,
    month INT NOT NULL,
    day INT NOT NULL,
    quarter INT NOT NULL
);

-- Таблиця фактів врожайності
CREATE TABLE Fact_Yield (
    id_area INT,          -- Зв'язок із таблицею Dim_Area
    id_seed INT,          -- Зв'язок із таблицею Dim_Seed
    id_date INT,          -- Зв'язок із таблицею DateDim
    yield DECIMAL(10, 2),   -- Врожайність
    PRIMARY KEY (id_area, id_seed, id_date), -- Складений ключ
    FOREIGN KEY (id_area) REFERENCES Dim_Area(id_area),
    FOREIGN KEY (id_seed) REFERENCES Dim_Seed(id_seed),
    FOREIGN KEY (id_date) REFERENCES DateDim(id_date)
);
```