КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій
Кафедра інтелектуальних технологій

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня «магістр»

НА ТЕМУ:

Інтелектуальна технологія прогнозування врожайності сільськогосподарських культур на основі супутникових даних

Галузь знань: 12 «Інформаційні технології»
Спеціальність: 122 «Комп'ютерні науки»
Освітньо-наукова програма «Технології штучного інтелекту»

Виконав:
студент 2 курсу магістратури
групи ТШІ-21
Горб Олександр Олександрович

Науковий керівник:
Сорока Петро Миколайович
кандидат фізико-математичних наук, доцент
(науковий ступінь, вчене звання)

Засвідчую, що в цій кваліфікаційній роботі немає запозичень з праць інших авторів без відповідних посилань

Студент(ка) _____  ___
підпис

Кваліфікаційна робота допущена до захисту
рішенням кафедри *інтелектуальних технологій*

Протокол № ____ від «____» травня 2022 р.

Зав. кафедри _____ доц. Іларіонов О.Є.
підпис

Київ 2022

РЕФЕРАТ

Кваліфікаційна робота складається зі вступу, 3 розділів, висновків, списку літератури з 51 джерела та 1 додатка. Загальний обсяг роботи 74 сторінки. Робота містить 6 таблиць і 33 рисунки.

Актуальність теми. Надзвичайно важливою для планування аграрного виробництва є проблема математичного моделювання та прогнозування врожайності сільськогосподарських культур. Водночас вплив природно-кліматичних, біологічних та організаційно-технологічних груп чинників надає вплив на результати такого планування, приводячи до високої похибки, що сягає більше 15%. Застосування побудови систем рівнянь економетрики, різного виду адаптивних моделей, а також сучасних методів нелінійної динаміки, здебільшого не приносять необхідних результатів, саме через це використання та дослідження нових моделей на основі методів машинного навчання в сільському господарстві, є надзвичайно актуальним, потрібним та перспективним напрямком.

Метою кваліфікаційної роботи є дослідити ефективність методів машинного навчання для прогнозування врожайності в сільському господарстві.

Об'єктом дослідження є прогноз врожайності сільськогосподарських культур методами машинного навчання.

Предметом дослідження є прогнозування врожайності сільськогосподарських культур на основі методів машинного навчання.

Результати роботи. У ході роботи були розглянуті методи машинного навчання, розроблено програмний додаток на основі них, який використовує моделі для прогнозування врожайності сільськогосподарських культур в Україні. Використовуючи створений додаток, проведено експериментальне прогнозування врожайності сільськогосподарських культур на 2021 рік.

Результатом експерименту була похибка прогнозу для кожного методу машинного навчання та кожної сільськогосподарської культури. На основі

аналізу отриманих результатів показано, що для обраних сільськогосподарських культур та вхідних даних – оптимальним варіантом стали моделі на основі нейронних мереж, що дали найбільш точний результат.

Ключові слова: методи машинного навчання, лінійна регресія, нейронні мережі, random forest, NDVI, VHI, прогнозування, врожайність, точне землеробство, сільськогосподарські культури.

ABSTRACT

Qualification work consists of an introduction, 3 chapters, conclusions, a list of references from 51 sources and 1 appendix. The total volume of the work is 74 pages. The work contains 6 tables and 33 figures.

Actuality of theme. The problem of mathematical modeling and forecasting of crop yields is extremely important for agricultural production planning. At the same time, the influence of natural-climatic, biological and organizational-technological groups of factors influences the results of such planning, leading to a high error of more than 15%. The use of econometric equation systems, various types of adaptive models, as well as modern methods of nonlinear dynamics, mostly do not bring the necessary results, because of this use and research of new models based on machine learning methods in agriculture is extremely relevant, necessary and promising direction.

The purpose of the qualification work is to investigate the effectiveness of machine learning methods for forecasting yields in agriculture.

The object of research is the forecast of crop yields by machine learning methods.

The subject of research is the forecasting of crop yields on the basis of machine learning methods.

Results of the work. In the course of the work the methods of machine learning were considered, a software application based on them was developed, which uses models for forecasting the yield of agricultural crops in Ukraine. Using the created application, experimental forecasting of crop yields for 2018 was conducted.

The result of the experiment was a forecast error for each machine learning method and each crop. Based on the analysis of the obtained results, it is shown that for the selected crops and input data - the best option were models based on neural networks, which gave the most accurate result.

Keywords: machine learning methods, linear regression, neural networks, random forest, NDVI, VHI, forecasting, yield, precision agriculture, crops.

ЗМІСТ

ВСТУП

Основним завданням науки та практики є отримання правдивих прогнозів про майбутню поведінку складних систем, об’єктів чи явищ на основі їхньої минулої поведінки. Прогнозування є важливою частиною прийняття сучасних інформаційно-технологічних рішень при проектуванні складних систем (паливно-енергетичних, сільських, інформаційно-комунікаційних тощо) та управлінні ними в невизначених умовах. Ефективність рішення оцінюється на основі подій і результатів, які відбуваються після його прийняття та виконання. Тому прогнозування та оцінка наслідків реалізації альтернативних виборів, зроблених на етапі їх формування та аналізу, дозволяє приймати кращі рішення та значно знижувати ризик несприятливих наслідків.

Багато проблем, які виникають на практиці, неможливо вирішити відомими раніше методами чи алгоритмами. Це тому, що ми не знаємо наперед механізми походження вихідних даних або відомої нам інформації недостатньо для побудови моделі прогнозу. Таким чином ми отримуємо дані з «чорного ящика». За цих умов нам не залишається нічого іншого, як вивчити послідовність доступних нам вихідних даних і спробувати побудувати прогнозну модель і в процесі її покращити. Підхід, який використовує попередні дані або зразки для першої побудови та покращення прогнозної моделі, називається машинним навчанням.

По-перше, машинні методи користуються великою популярністю через їхню високу ефективність і здатність обробляти величезні обсяги інформації. Машинне навчання – надзвичайно широка наукова галузь, яка динамічно розвивається та використовує різноманітні теоретичні та практичні методи.

Машинне навчання сьогодні використовується практично у всьому, від звичайного запиту пошукової системи до безпілотних автономних транспортних засобів. Здається, що традиційна галузь, наприклад сільське

господарство, не є винятком. З року в рік у сільському господарстві широко використовуються методи прогнозування та підвищення врожайності на основі машинного навчання. Саме таким методам і присвячена дана робота. Сьогодні такими технологіями користуються як великі компанії, так і холдинги та невеликі стартапи.

Проблема математичного моделювання та точного прогнозування врожайності сільськогосподарських культур є важливою для планування сільськогосподарського виробництва. При цьому вплив ряду груп природно-кліматичних, біологічних та організаційних факторів має різноспрямований вплив на результати прогнозу, що призводить до неприпустимо високої похибки понад 15%. Застосування класичних підходів до математичного моделювання, таких як побудова економетричних систем рівнянь, різних типів адаптивних моделей, а також сучасних методів нелінійної динаміки, не завжди приводить до адекватних результатів. Тому використання та дослідження нових класів математичних моделей у сільському господарстві, таких як моделі на основі методів машинного навчання, є актуальним та перспективним напрямком.

У цій кваліфікаційній роботі ми познайомимося з деякими сучасними математичними задачами цієї галузі та їх вирішенням, основною проблемою яких є побудова моделей та оцінка якості їх прогнозів.

Метою кваліфікаційної роботи є дослідження ефективності методів машинного навчання у прогнозуванні врожайності сільськогосподарських культур.

Об’єктом дослідження є прогнозування врожайності сільськогосподарських культур за допомогою методів машинного навчання.

Предметом дослідження є системи прогнозування врожайності сільськогосподарських культур на основі різних методів машинного навчання.

Завданнями дослідження є:

1. аналізувати літературні та інші джерела, зокрема Інтернет, для прогнозування врожайності;
2. описати концепцію точного землеробства та його наслідки для підвищення врожайності сільськогосподарських культур;
3. розгляд різних підходів і методів машинного навчання для побудови систем прогнозування;
4. вибір та опис вхідних даних та аналіз;
5. вибрати необхідну архітектуру для кожної моделі машинного навчання та обґрунтувати свій вибір;
6. розробка системи прогнозування врожайності сільськогосподарських культур на основі різних методів машинного навчання;
7. описати та проаналізувати отримані результати;
8. визначити адекватність і точність прогнозування отриманих систем.

РОЗДІЛ 1 АНАЛІЗ ПРОБЛЕМИ ВИКОРИСТАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ КУЛЬТУР ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Аналіз публікацій та сучасного стану проблеми

Останні кілька десятиліть ознаменувалися проривом у галузі обчислювальної техніки та інформаційних технологій. Завдяки цьому обробляються величезні обсяги даних у різних галузях, таких як медицина, біологія, фінанси та маркетинг. Завдання розуміння цих даних призвело до появи нових інструментів у галузі статистики та до виникнення нових галузей, таких як дейта майнінг, машинне навчання та біоінформатика.

Сьогодні машинне навчання стає все популярнішим і поступово входить у всі аспекти повсякденного життя. Провідні виробники зробили машинне навчання невід'ємною частиною смартфонів, мобільних програм та побутової техніки, щоб забезпечити їх ефективніше використання. Зараз практично немає такої проблеми, яку б не намагалися вирішити за допомогою машинного навчання: від пошуку помилок у тексті до створення творів мистецтва [1].

Методи машинного навчання завоювали популярність завдяки своїй високій ефективності та здатності обробляти великий обсяг доступної в даний час інформації. Наприклад, сучасні системи розпізнавання зображень досягають точності 93,9% [2].

Схоже, що така традиційна сфера, як сільське господарство, не є винятком і останні роки активно використовує машинне навчання. Поки автовиробники тестують можливості автопілота, виробники тракторів уже кілька років продають автономні трактори [3].

У сучасних умовах ринкових відносин та незалежності сільськогосподарських земель питання прогнозування врожайності стає дедалі актуальнішим.

Серед багатьох показників, що характеризують діяльність сільського господарства, особливу увагу слід приділити техніко-економічним показникам, а також показникам урожайності, які є комплексними показниками. З одного боку, ці показники є цінною інформацією для прогнозування, планування та прийняття управлінських рішень, з іншого боку, вони можуть бути використані для оцінки ефективності сільського господарства [4].

Урожайність сільськогосподарських культур є складним показником з погляду моделювання, оскільки формування рослин залежить не тільки від виробничих чинників, а й від біологічних і кліматичних показників. Тому показники, що впливають на врожайність, можна розділити на дві групи: технологічний рівень землеробства та погодні фактори [5].

Отримання точного прогнозу врожайності сільськогосподарських культур дозволяє правильно підійти до формування резервних фондів продовольства та побудувати ефективну та обґрунтовану зовнішньоторговельну політику [4].

Це робить пріоритетною методологію побудови систем комплексного агрометеорологічного обслуговування сільського господарства, яка має єдину методологічну основу для всіх компонентів системи як для набору сільськогосподарських культур, так і для методів прогнозування та оцінки умов формування рослин. Такою методологічною основою є математичні моделі [6].

Математичні моделі — це опис предмета чи явища на основі математичного апарату (рис. 1.1). Вони використовуються для вивчення систем і впливу різних параметрів на їх поведінку. Математичні моделі особливо корисні, коли важко експериментувати з реальним об’єктом або явищем, наприклад, коли явище рідкісне або експеримент занадто дорогий.

Культура	Модель врожайності	R^2
Озима пшениця	$Y_O = 0{,}67 \cdot T^2 - 0{,}68 \cdot T - 0{,}011 \cdot R^2 - 6{,}82 \cdot R + 0{,}052 \cdot X + 61{,}34 \cdot D^2 + 0{,}001 \cdot D + 0{,}03 \cdot Q^2 + 0{,}0014 \cdot \Sigma D^2 - 0{,}00253$	0,892
Кукурудза на зерно	$Y_K = -3 \cdot 10^{-5} \cdot \Sigma D^2 - 74 \cdot 10^{-5} \Sigma D - 33 \cdot 10^{-4} R^2 - 4{,}29 \cdot R + 37 \cdot 10^{-5} X^2 + 0{,}27 \cdot X + 0{,}00124 \cdot W^2 - 0{,}237 \cdot W - 0{,}0001 \cdot Q^2 + 1{,}855 \cdot Q + 0{,}063$	0,747
Люцерна на корм	$Y_Л = -0{,}0011 \cdot W^2 - 311 \cdot K^3 + 736 \cdot K + 0{,}091 \cdot B^2 - 10{,}86 \cdot B + 0{,}96 + 0{,}063 \cdot Q^2 - 10{,}86 \cdot Q - 3{,}11 \cdot D^2 + 74{,}56 \cdot D + 0{,}0026 \cdot X^2 + 0{,}063 \cdot X$	0,872

Рисунок 1.1 Приклад математичного моделювання врожайності сільськогосподарських культур [7]

В галузі математичного моделювання врожайності сільськогосподарських культур досягнуто значного прогресу: розроблено моделі фундаментальних життєвих процесів, що впливають на продуктивність рослинності, розроблено моделі формування врожайності для прогнозування, планування та управління в сільському господарстві [6].

В останні кілька десятиліть методи машинного навчання, які є своєрідною еволюцією математичного моделювання, стають все більш популярними для моделювання врожайності сільськогосподарських культур.

В основному це пов'язано зі зростанням обчислювальних потужностей та їх доступністю. Крім того, значно зросла кількість доступних наборів даних. За прогнозами, до 2025 року їхня кількість збільшиться більш ніж у чотири рази порівняно з 2021 роком. (рис. 1.2) [8].

Слід зазначити значне збільшення доступних супутникових даних, що призвело до прискореного розвитку методів та інформаційних технологій прогнозування врожайності з урахуванням супутникової інформації останніми роками [9].

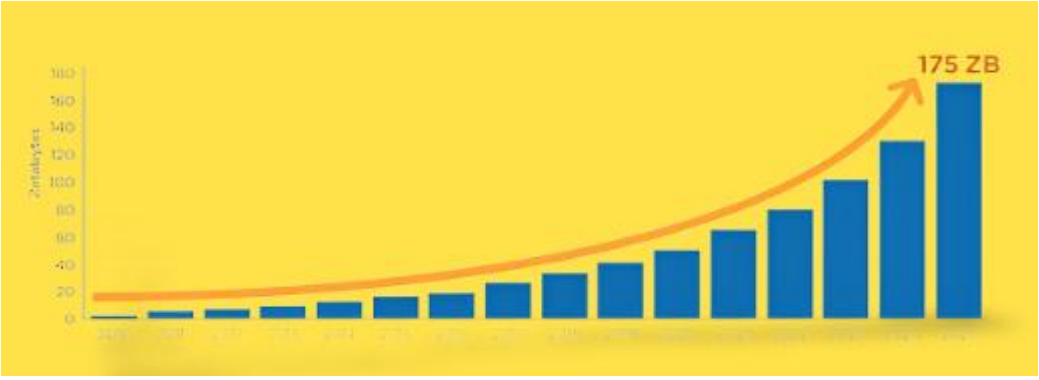

Рисунок 1.2 Кількість даних у світі з прогнозом до 2025 року [8]

Серед вітчизняних та зарубіжних вчених є значна кількість наукових праць, пов'язаних із використанням методів машинного навчання для прогнозування врожайності сільськогосподарських культур.

Наприклад, у роботі Т.Ф. Бекмуратової, Д.Т. Мухамедієвої, О.Ж. Бобомурадова «Нечітка модель прогнозування врожайності» [10] використовувалися нечіткі множинні моделі типу Сугено для прогнозування врожайності бавовнику різних сортів за нечітко заданих типів ґрунту, режимів зрошення та внесення добрив, а також погодних умов.

У роботі було використано аналітичне моделювання прогнозу врожайності та проведено аналіз аналітичних залежностей прогнозу врожайності залежно від різних факторів. Як вхідні параметри використовувалися такі нечіткі дані: погода під час посіву, водопостачання, погода під час вегетаційного періоду, погода під час збирання, тип ґрунту, тип сорту, тип внесення добрив. Результати експерименту показали високу ефективність прогнозування, зокрема, на основі моделей Сугено.

У роботі А. А. Темірової «Алгоритм лінійного клітинного автомату для прогнозування врожайності зернових» [11] для побудови моделі прогнозування врожайності зернових на основі часових рядів використано алгоритм лінійного клітинного автомату. Використовували моделі прогнозування, засновані на клітинних автоматах різної конфігурації. Для моделювання структури клітинного автомату та її навчання використовувався генетичний алгоритм. Як вхідні дані використовувався

часовий ряд річної врожайності зернових з 1896 по 2021 рік. У ході роботи було підтверджено придатність моделі прогнозування клітинного автомата для прогнозування числового ряду врожайності зернових. Помилка прогнозування становила близько 8,4%.

У роботі Н.М. Куссуля, А.В. Колотія, С.В. Яцківа, Т.В. Олійника «Регресійні моделі прогнозування врожайності зернових в Україні за супутниковими даними різної природи» [12] використано регресійні моделі прогнозування врожайності зернових за супутниковими даними. У статті порівнюється використання різних супутникових даних для прогнозування врожайності пшениці озимої в Україні для різних регіонів. Аналіз результатів показав, що використання індексу NDVI призводить до більшої помилки прогнозу, ніж FAPAR і VHI, але обмежена кількість даних не говорить про те, що NDVI завжди даватиме менш точний прогноз. Тому автори вважають за доцільне не обмежуватися рамками однієї моделі прогнозування.

У роботі Н.Д. Заводчикової, Н.В. Впешилова, С.С. Таспаєва «Використання нейромережевих технологій у прогнозуванні ефективності виробництва зерна» [13] використовуються імітаційні моделі на основі нейронної мережі для прогнозування врожайності зерна та розраховується врожайність при зміні деяких параметрів на основі отриманої моделі. Як вихідні дані використовувалися 9 показників, які, згідно з кореляційним аналізом, надавали найбільший вплив на врожайність. Отримані результати підтвердили доцільність використання нейронних мереж на вирішення завдання оптимізації виробництва зерна.

У роботі Л.А. Хворої, Н.В. Гавриловської «Прогнозування врожайності зернових культур: методи та розрахунки» [14] рік визначався рік-аналог за аналогією з використанням методів класифікації та розпізнавання образів для прогнозування врожайності ярої пшениці. Для створення моделі роки кластеризуються за певними параметрами, а для створення прогнозу рік класифікується на основі кластерів, отриманих у моделі.

На основі дисперсійного аналізу для всіх показників було обрано такі параметри: сума ефективних температур, кількість опадів, кількість днів з опадами, дефіцит вологи насичення. В результаті було розроблено модель, за допомогою якої можна взяти погодний сценарій всього вегетаційного періоду року-аналогу та створити оновлений прогноз урожайності на поточний рік.

Використовуючи дані по рокам-аналогам, попередній прогноз можна зробити через два-три тижні вегетаційного періоду. Автори зазначають, що класифікацію та прогнозування врожайності ярої пшениці слід розглядати як початковий етап роботи з оцінки врожайності зернових. Це показує, що цей ключовий показник має певну суть і дає добрий прогноз.

У роботі А.Г. Гагаріна А.Ф. Рогачова «Прогнозування врожайності на основі аналізу крос-регіональних даних» [15] використовували нейронні мережі для побудови моделей прогнозування врожайності озимої пшениці. Для навчання використано вибірку з 54 динаміки зі значеннями врожайності озимої пшениці за 21 рік.

На підставі проведених прогнозних експериментів та аналізу статистичних характеристик часових рядів урожайності досліджуваних культур було обґрунтовано вибір програмного забезпечення, структури штучної нейронної мережі, її навчання та можливість короткострокового прогнозування з помилкою 15-20%.

Аналізуючи ці і подібні роботи, можна побачити, що у них застосовуються різні способи машинного навчання, типи вхідних даних, сільськогосподарські культури, терміни прогнозування тощо. З цього можна зробити висновок, що має місце значний простір для досліджень у цій галузі і бракує систематичних досліджень щодо ефективності різних параметрів.

1.2 Аналіз існуючих систем прогнозування врожайності

Незважаючи на велику кількість наукових праць на цю тему з досить

точними результатами, нині немає окремих комерційних систем прогнозування врожайності. Проте останнім часом широкого поширення набула концепція точного землеробства.

Точне землеробство - це концепція використання інтегрованих систем, заснованих на різних технологічних рішеннях, які можуть підвищити врожайність та сприяти кращому управлінню сільськогосподарськими ресурсами.

Для точного землеробства використовується низка технологій та показників:

- географічні інформаційні системи (ГІС);
- системи глобального позиціонування (GPS);
- дистанційне зондування землі (ДЗЗ);
- технологія змінної норми висіву;
- дані зі спеціальних датчиків тощо.

Концепція точного землеробства полягає в тому, що умови для розвитку культур у різних частинах одного й того ж самого поля подібні, але неднакові, тобто у межах одного поля існують неоднорідності (рис. 1.3).

Рисунок 1.3 Приклад неоднорідності на одному полі

Точне землеробство засноване на використанні найбільш детальних карт конкретних полів. Наявні кадастрові карти дають мало корисної інформації, а саме межі полів. На додаток до цієї інформації потрібні дані про вологість ґрунту, хімічний склад, кут нахилу поверхні, сонячну радіацію тощо. Чим більше факторів містить така карта, тим точніше працюють комп'ютерні та супутникові технології, тим швидше та ефективніше можна вносити корективи у виробництво.

Системи точного землеробства часто є системами підтримки прийняття рішень, тому основі створюваних ними карт складаються точні рекомендації. Для кожної ділянки розраховується необхідна кількість води, насіння та добрив.

На основі рекомендацій складаються інструкції, які завантажуються до бортового комп'ютера сільськогосподарської техніки. Виконуючи інструкції, людина лише перевіряє правильність виконання машиною. Техніка на основі супутникової навігації переміщається по полю та **в**носить насіння та добрива, регулюючи їх кількість на задану площу поля відповідно до отриманих інструкцій. Завдяки використанню GPS шлях прокладається таким чином, що між оброблюваними маршрутами немає прошарків або розривів.

Таким чином, використовуючи точне землеробство, можна досягти таких переваг:

- значно знижуються витрати на насіння та матеріали і, як наслідок, собівартість продукції.

- збільшується врожайність та прибуток;

- отримується продукція найвищої якості;

- покращується якість ґрунту;

- знижується негативний вплив на навколишнє середовище за рахунок зменшення кількості добрив, що вносяться.

Аналізуючи системи точного землеробства, можна отримати такі параметри поля:

- просторові дані;

- карти врожайності;
- фотографії та карти NDVI;
- дані польових пристроїв;
- стан ґрунту;
- дані про внесені добрива;
- метеорологічні дані;
- тощо.

На основі цього набору даних можна побудувати точні моделі для прогнозування врожайності кожного окремого поля. Крім того, більшість систем точного землеробства мають архіви зі значеннями протягом останніх десяти років, що ще більше спрощує завдання. Проте з невідомих причин більшість цих систем все ще не можуть прогнозувати врожайність. Тому, з огляду на актуальність цієї теми, було б доцільно додати можливість прогнозування врожайності у системах точного землеробства.

1.3 Постановка задачі

У перерахованих вище роботах використовуються різні методи для створення систем прогнозування. Деякі розглядають вплив різних вхідних даних на продуктивність системи, але порівнюються моделі, засновані на різних методах машинного навчання.

Г.С. Розенберг та інші [16] відзначають, що особливість прогнозування на сучасному етапі полягає, перш за все, у тому, що одне й те саме явище можна побачити, використовуючи декілька різних і більш-менш еквівалентних моделей (прояв принципу множинності моделей). На їх думку, основним недоліком існуючих систем прогнозування є те, що передбачення конкретного часового ряду ґрунтується лише на одному алгоритмі. Іншими словами, справжній механізм генерації цього ряду передбачається єдиним і що він добре апроксимується одним із алгоритмів.

Тому залишається невирішеним питання: система, заснована на якомусь методі, дасть більш точний результат.

Це питання надзвичайно важливе, оскільки відмінності в точності прогнозу, навіть у відсотках, можуть суттєво вплинути на кількість зібраного врожаю, а отже, на доходи фермерів та країни. Тому в цій роботі розглядається побудова системи прогнозування врожайності сільськогосподарських культур з використанням різних методів машинного навчання та порівняння їх ефективності.

РОЗДІЛ 2 ОПИС ТА АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ, ВХІДНИХ ДАНИХ ТА АРХІТЕКТУРИ МОДЕЛЕЙ

2.1 Аналіз методів машинного навчання

2.1.1 Класифікація методів машинного навчання

Машинне навчання - це область штучного інтелекту, що використовує алгоритми та статистичні моделі, які застосовують комп'ютерні системи для ефективного виконання конкретних завдань без використання чітких інструкцій, покладаючись на шаблони та логічні висновки. Алгоритми машинного навчання будують математичні моделі на основі вибірок даних, щоб робити передбачення чи умовиводи для досягнення конкретної цілі без явного програмування [17, 18].

Основна мета навчальної системи - робити висновки (узагальнення) з урахуванням власного досвіду [18]. Тут узагальнення - це здатність даної системи точно виконувати нові, ще невирішені завдання після отримання досвіду на базі навчальної вибірки. На її основі система має побудувати загальну модель, за допомогою якої вона робитиме передбачення у нових випадках [19].

Термін «машинне навчання» запроваджено 1959 року Самюелем Артуром. Ще на початку розвитку штучного інтелекту деякі вчені цікавилися машинами, які навчаються на основі даних. Вони випробовували різні символьні методи задля досягнення цієї мети. В основному це були персептрони та інші моделі, засновані на узагальнених лінійних статистичних моделях, які пізніше назвали "нейронними мережами" [20].

На початку 1980-х років експертні системи стали домінувати у сфері штучного інтелекту [21]. Це, а також брак обчислювальних потужностей та проблеми з накопиченням даних призвели до занепаду машинного навчання (яке ще не було відокремлено від штучного інтелекту) [22].

У 1990-х роках машинне навчання відродилося, виділившись в окрему підгалузь штучного інтелекту. Основною метою нової підгалузі стало вирішення практичних завдань. Вона почала грунтуватися на статистичних методах і методах теорії ймовірностей, на відміну символьних методів, які використовувались у штучному інтелекті [21]. Крім того, значно зросли обчислювальні потужності, а розвиток Інтернету полегшив доступ до даних, що, своєю чергою, сприяло розвитку машинного навчання.

У машинному навчанні існує багато методів і алгоритмів. Їх загальна класифікація наведена на рис. 2.1.

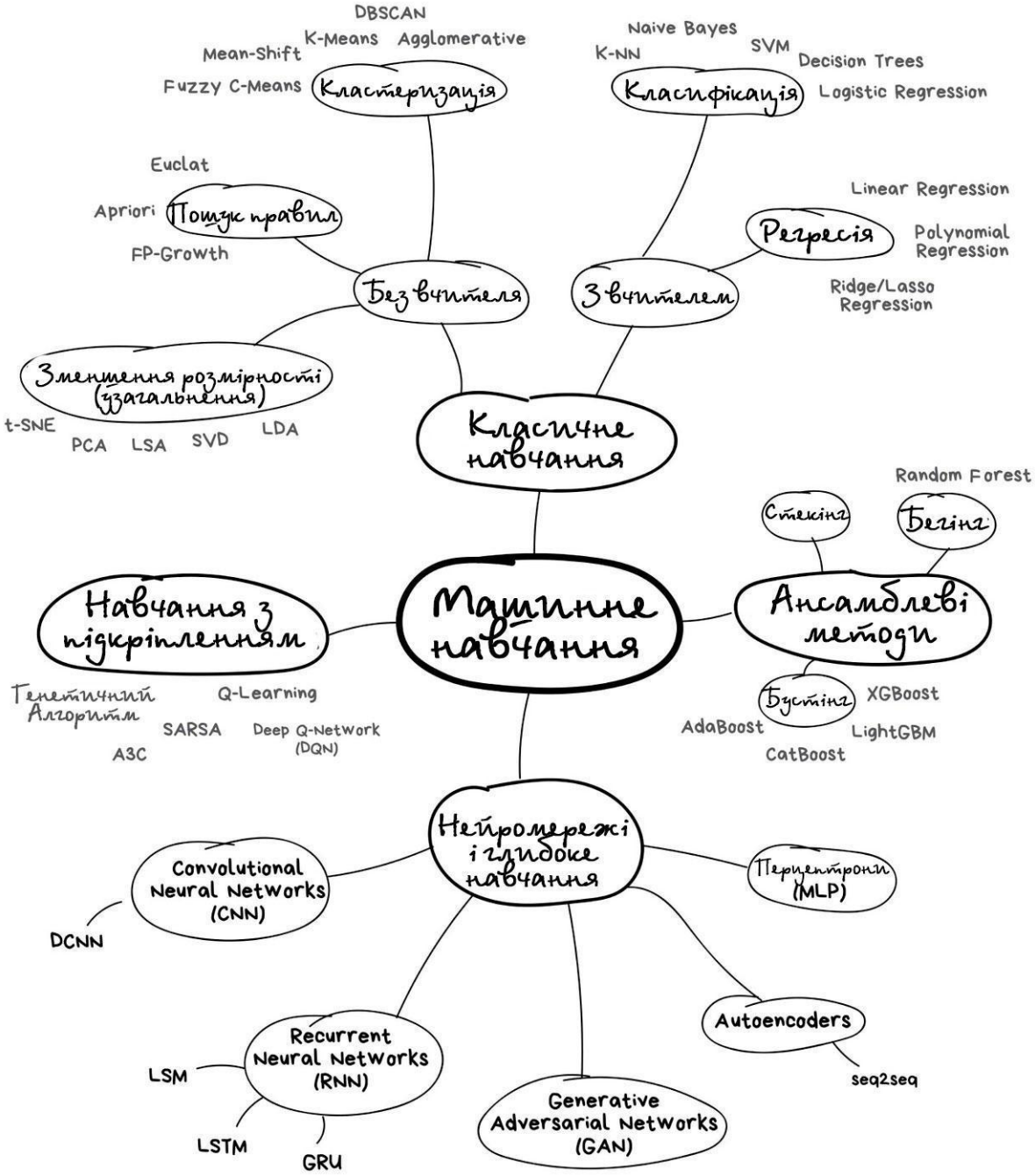

Рисунок 2.1 Загальна класифікація алгоритмів машинного навчання [23]

Існують різні типи машинного навчання: навчання з учителем, навчання без вчителя та навчання з підкріпленням. Давайте докладніше розглянемо кожен із них та проблеми, які вони вирішують.

Навчання з учителем - це метод машинного навчання, що передбачає маркування всіх даних. У процесі навчання результат моделі порівнюється з правильною відповіддю, і на цьому ґрунтується навчання. Цей метод машинного навчання вирішує такі проблеми:

- проблеми регресії - отримання скаляра або вектора на основі вхідних даних;

- проблеми класифікації – на основі вхідних даних визначити, до якого класу належить об'єкт.

Навчання без вчителя - це метод машинного навчання, у якому дані не маркуються. Навчаючи модель, вони повинні знаходити закономірності та робити висновки на основі даних. Цей метод вирішує такі завдання:

- кластеризація – пошук подібних об'єктів, об'єднання їх у класи, причому класи визначаються самостійно;

- зниження розмірності даних – зменшення розмірності даних шляхом збору певних атрибутів на високому рівні абстракції;

- пошук правил (асоціацій) - пошук закономірностей чи правил на базі даних.

Навчання з підкріпленням - це метод машинного навчання, у якому система (агент) взаємодіє з певним середовищем. Мета такого навчання – мінімізувати помилки, а не запам'ятати чи прорахувати усі можливі варіанти. Прикладами застосування є автопілот чи штучний інтелект у іграх.

Таким чином, для розв’язання поставленої задачі підходить навчання з вчителем, оскільки необхідно розв’язати саме задачу регресії: передбачити числовий результат (вихід) на основі вхідних даних. Розглянемо відповідні методи машинного навчання, придатні для вирішення цього завдання.

2.1.2 Лінійна та поліноміальна регресія

Лінійна регресія – це лінійний підхід до моделювання взаємозв'язків між скалярними залежними змінними та векторами незалежних змінних

(рис. 2.2). Із однією залежною змінною вона називається простою лінійною регресією. З декількома залежними змінними – множинною лінійною регресією [24]. Як випливає з назви, метод вирішує регресійні завдання.

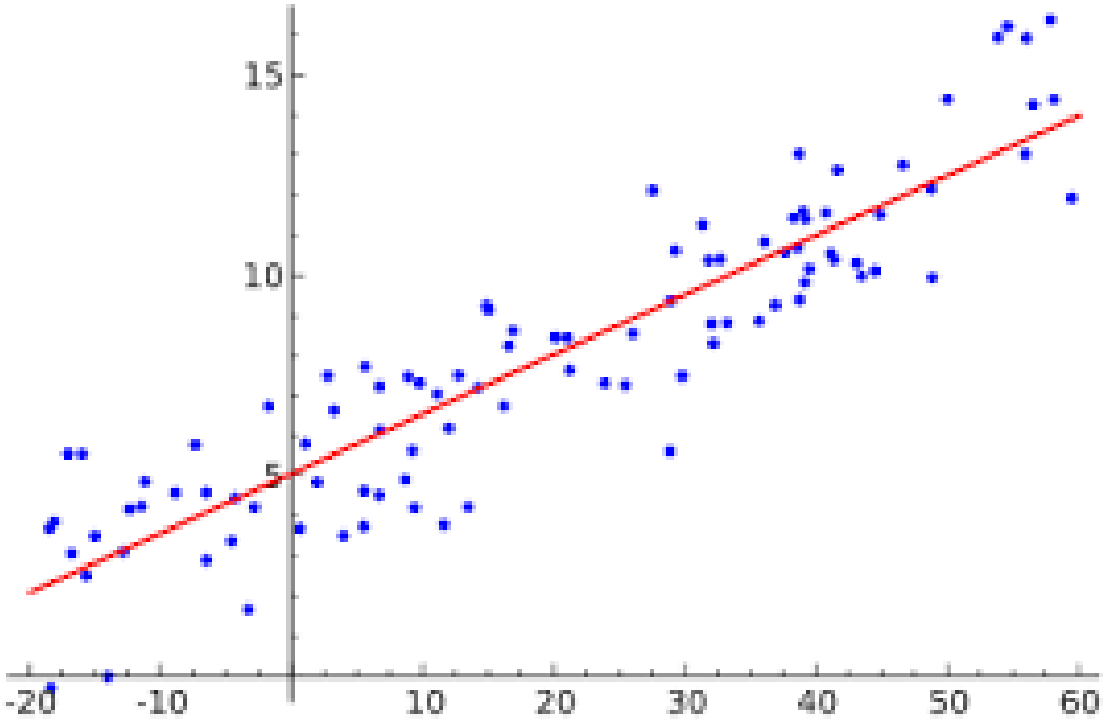

Рисунок 2.2 Приклад лінійної регресії [25]

У лінійній регресії зв'язки моделюються з використанням функції лінійного прогнозування, параметри якого невідомі і обчислюються на основі даних. Такі моделі називають лінійними [26].

У загальному випадку модель лінійної регресії виглядає наступним чином:

$$y_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \varepsilon_i y_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \varepsilon_i,$$

де

$i = 1, \ldots, n$ – номер випадку (порядковий номер експерименту);

y_i – залежна (передбачувана) змінна (у i-тому випадку);

x_{ip} – p-та координата вектора незалежних змінних (у i-тому випадку);

$\theta\ \theta_p$ – вектор параметрів;

ε_i – непередбачена випадкова похибка, яка додає «шум» (у i-му випадку).

Для зручності, усі *n* вирази можна представити у матричній формі запису:

$$\boldsymbol{y} = X\theta + \varepsilon \; \boldsymbol{y} = X\theta + \varepsilon \; ,$$

де

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$ – вектор-стовпчик залежної змінної (усіх *i*-их випадків);

$$X = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} X = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$ – матриця векторів незалежної змінної (усіх *i*-их випадків);

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$$ – вектор-стовпчик параметрів;

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$ – вектор-стовпчик випадкових похибок (усіх *i*-их випадків) [24].

Якщо дані мають нелінійний зв'язок то слід використовувати поліноміальну регресію (рис. 2.3). Відрізняється від лінійної регресії використанням поліному вищої степені (степінь більше 1).

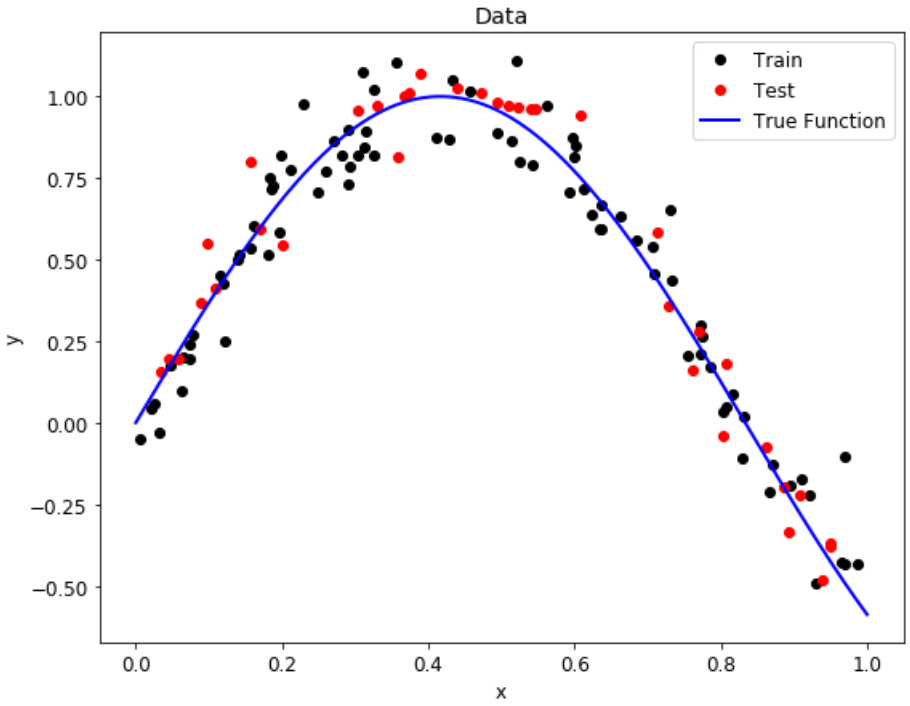

Рисунок 2.3 Приклад поліноміальної регресії [25]

У поліноміальній регресії необхідно використовувати поліном такої міри, щоб він краще апроксимував необхідну залежність. Для пошуку оптимальної степені використається крос-валідація.

Якщо використовувати поліном занадто низької степені, модель стає дуже простою (недостатньою) і не може правильно знайти обмеження (рис. 2.4). Якщо модель дуже проста, помилка буде високою як у навчальній, так і в тестовій вибірці.

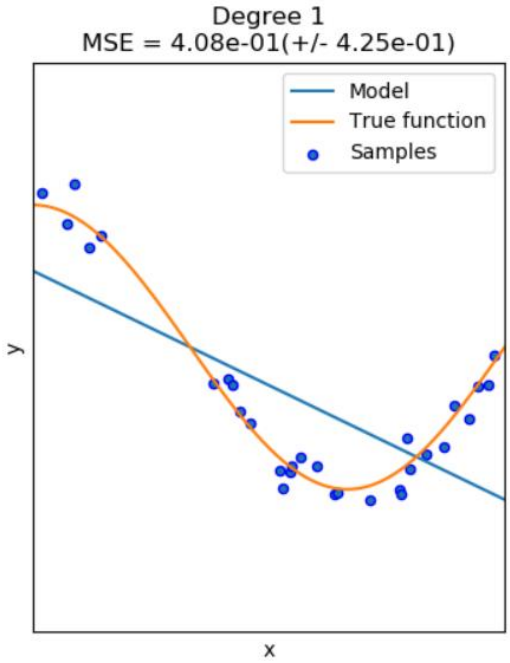

Рисунок 2.4 Занадто проста поліноміальна модель (underfitting) [27]

Якщо поліном використовує високу степінь, модель перенавчається (перенавчання) і намагається "запам'ятати" якнайбільше прикладів з навчальної вибірки (рис. 2.5). При перенавченні помилка на навчальній вибірці мінімальна, але велика на тестовій.

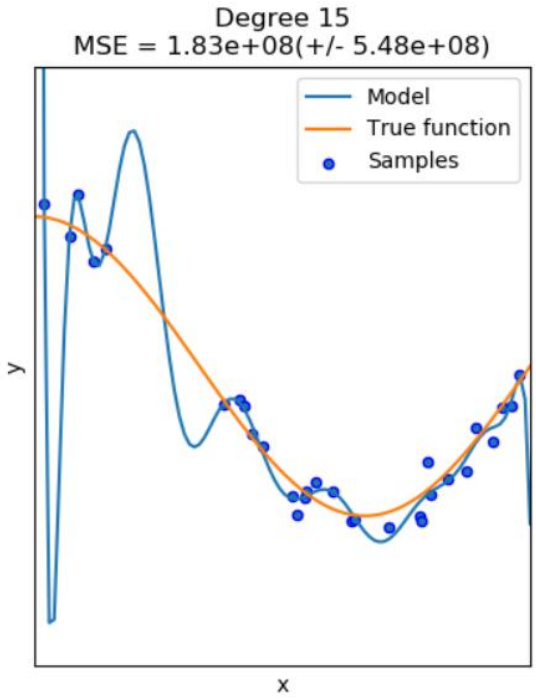

Рисунок 2.5 Занадто складна, перенавчена модель (overfitting) [27]

Для знаходження параметрів $\theta\ \theta$ існує декілька методів. Суть методів полягає у тому, аби знайти такі $\theta\ \theta$, які б мінімізували суму квадратів різниці між реальним та передбачуваним значенням у всіх випадках. Таким чином функція втрат виглядає таким чином:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2,$$

де $m\ m$ – кількість випадків, $y^{(i)}y^{(i)}$ – реальне значення у *i*-му випадку, $h_\theta\left(x^{(i)}\right)h_\theta\left(x^{(i)}\right)$ – прогнозоване значення для *i*-го випадку.

Найпопулярнішим методом є градієнтний спуск. Суть методу:

Повторювати доки помилка > ε або досягнута певна кількість ітерацій:

Одночасно змінювати усі ваги за формулою:

$$\theta_j = \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) * x_j \theta_j = \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) * x_j,$$

де $\alpha\ \alpha$ – параметр швидкості навчання.

Або у матричній формі:

$$\theta = \theta - \frac{\alpha}{m}X^T(h_\theta(x) - y)\theta = \theta - \frac{\alpha}{m}X^T(h_\theta(x) - y).$$

Однак цей метод має свої недоліки:

- необхідно вибрати $\alpha\ \alpha$;
- пошук локальних мінімумів;
- при великій різниці між значеннями змінних навчання буде повільніше (необхідно використовувати нормалізацію або інші методи попередньої обробки).

Ще один метод пошуку $\theta\ \theta$ називається нормальним рівнянням і описується такою формулою:

$$\theta = (X^T X)^{-1}X^T y\ \theta = (X^T X)^{-1}X^T y.$$

Цей метод має наступні переваги:

- не вимагає повторень;
- дає максимально можливу точність;

- швидкість не залежить від різниці між значеннями змінної.

Однак у нього є й свої недоліки:

- не завжди можна знайти обернену матрицю і тому цей метод не завжди можна застосувати;
- при досить великій кількості змінних він працює повільно через складність алгоритму пошуку оберненої матриці $O(n^3)O(n^3)$.

2.1.3 Нейронні мережі

Штучна нейронна мережа - це метод машинного навчання, що є мережею штучних нейронів, які отримують на вхід сигнал, на підставі якого вони змінюють свій внутрішній стан і видають початкові значення (залежно від внутрішнього стану та вхідних даних) [28].

Штучні нейрони ґрунтуються на їхніх біологічних аналогах. Тому ще спочатку історії їх створення виділяли 2 напрями. Перший орієнтувався на моделювання біологічних процесів, другий – на їхнє практичне застосування у задачах [29].

У загальному випадку нейрон та його робота виглядає наступним чином (рис. 2.6):

1) на вхід кожного нейрона подається сигнал (x_j) (може надходити від попередників нейрона *i*);

2) вхідний сигнал множиться на відповідну вагу (w_{ij});

3) отримані значення складаються;

4) отримана сума подається на функцію активації, що формує вихідне значення нейрона.

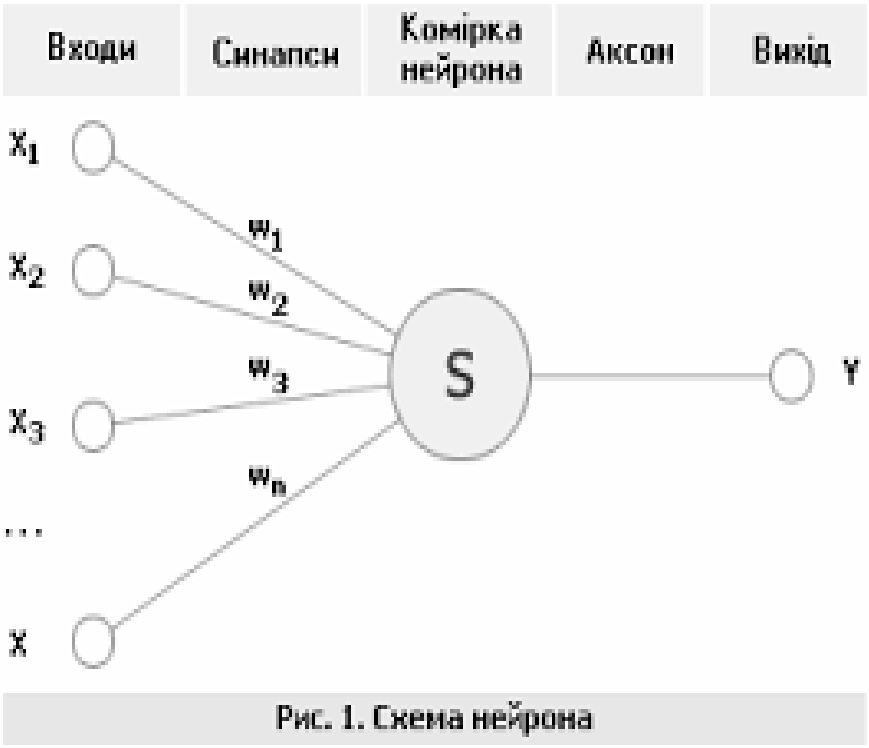

Рисунок 2.6 Ілюстрація штучного нейрона [30]

Нелінійні функції були обрані як функція активації тому, що вони дозволяють вирішувати нетривіальні завдання. Наприклад, згідно з універсальною теоремою апроксимації (теорема Цибенко), нейронна мережа прямого поширення із прихованим шаром із сигмоїдною функцією активації може апроксимувати будь-яку безперервну функцію багатьох змінних з довільною точністю [31]. Якщо використовуємо лінійні функції, то нейронні мережі вирішують дуже обмежений клас завдань, які можуть вирішувати функції лінійної апроксимації. Функції активації наведено у табл. 2.1.

Таблиця 2.1 Функції активації

Назва	Рівняння	Область значень
Сигмоїда	$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$	(0, 1)
Гіперболічний тангенс	$f(x) = \tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$	(-1, 1)
Функція Гауса	$f(x) = e^{-x^2}$	(0, 1]

(дзвоноподібна)		
ReLU (напівлінійна)	$f(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases}$	$[0, \infty)$
Напівлінійна з насиченням	$f(x) = \begin{cases} 0, x \leq 0 \\ x, 0 < x < 1 \\ 1, x \geq 1 \end{cases}$	[0, 1]
Порогова	$f(x) = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases}$	{0, 1}
Трикутна	$f(x) = \begin{cases} 1 - \lvert x \rvert, \lvert x \rvert \leq 1 \\ 0, \lvert x \rvert > 1 \end{cases}$	[0, 1]

Ознайомимося з найпопулярнішими функціями активації.

Найпопулярнішою функцією активації є сигмовидна (рис. 2.7). Це нелінійна, плавна й монотонно зростаюча функція.

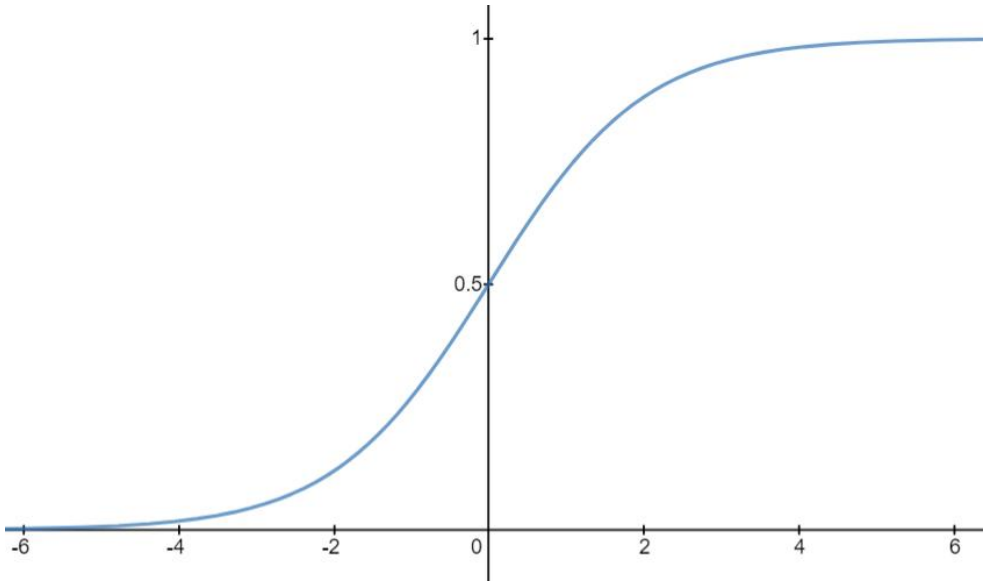

Рисунок 2.7 Графік сигмоїдної функції

Він має наступні переваги:

- має діапазон значень (0,1), який нормалізує вихідне значення для кожного нейрона;
- має гладкий градієнт, який запобігає різким змінам (стрибкам) при обчисленні значень;

- у діапазоні x x від -2 до 2 значення y y швидко змінюється й має тенденцію до наближення до одної з асимптот, що дозволяє робити чіткі передбачення класу.

Однак у нього є один недолік – при наближенні до асимптоти значення похідної сильно зменшується, що негативно позначається на швидкості навчання.

Гіперболічний тангенс подібний на сигмоїду як за виглядом (рис. 2.8), так і за деякими властивостями. Проте має відмінності й свої особливості.

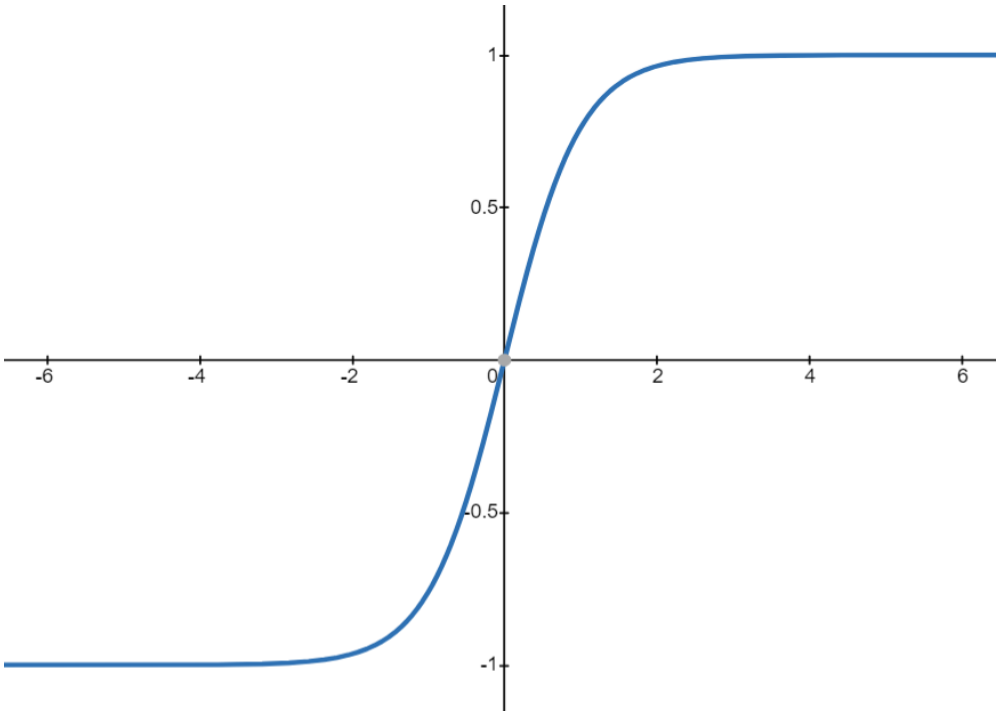

Рисунок 2.8 Графік гіперболічного тангенсу

Переваги:

- усі переваги сигмоїди, крім нормалізації;
- має негативні значення, що може бути корисно при роботі з ними;
- порівняно з сигмоїдою швидше сходиться, за рахунок вищого значення похідної біля нуля.

Недоліки:

- відсутня нормалізація;
- гірше працює у задачах класифікації.

Ще однією популярною функцією є *ReLu* (Rectified linear unit). Не дивлячись на її візуальну схожість на лінійну функцію (рис. 2.9), насправді є нелінійною.

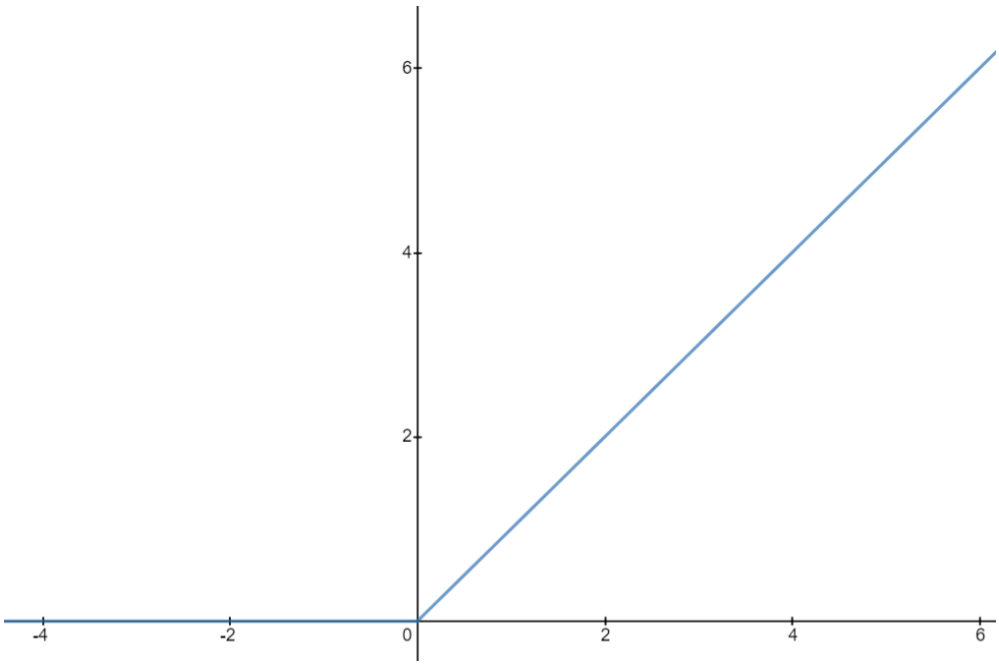

Рисунок 2.9 Графік ReLu

Переваги ReLu:

- похідна обчислюється швидко і легко;
- спрощує нейронну мережу, оскільки не всі нейрони активуються, що позитивно впливає на швидкість навчання.

Однак у ReLu є мінус: частина функції має похідну 0, тому деякі ваги не змінюються під час тренування.

На питання, яку функцію активації краще вибрати, не можна відповісти однозначно. Зазвичай функції вибираються для даної задачі на основі їх характеристик (наприклад, сигмоїда краще підходить для задачі класифікації, а гіперболічний тангенс - гірше) або через їхню подібність до апроксимованих функцій.

Нейронна мережа формується шляхом з'єднання виходів одних нейронів із входами інших, у результаті виходить зважений орієнтований граф. Залежно від з'єднання нейронів, формується архітектура мережі. Нині є багато різних архітектур. Вони показані на рис. 2.10. Тип архітектури

нейронної мережі вибирається залежно від завдання та необхідних властивостей.

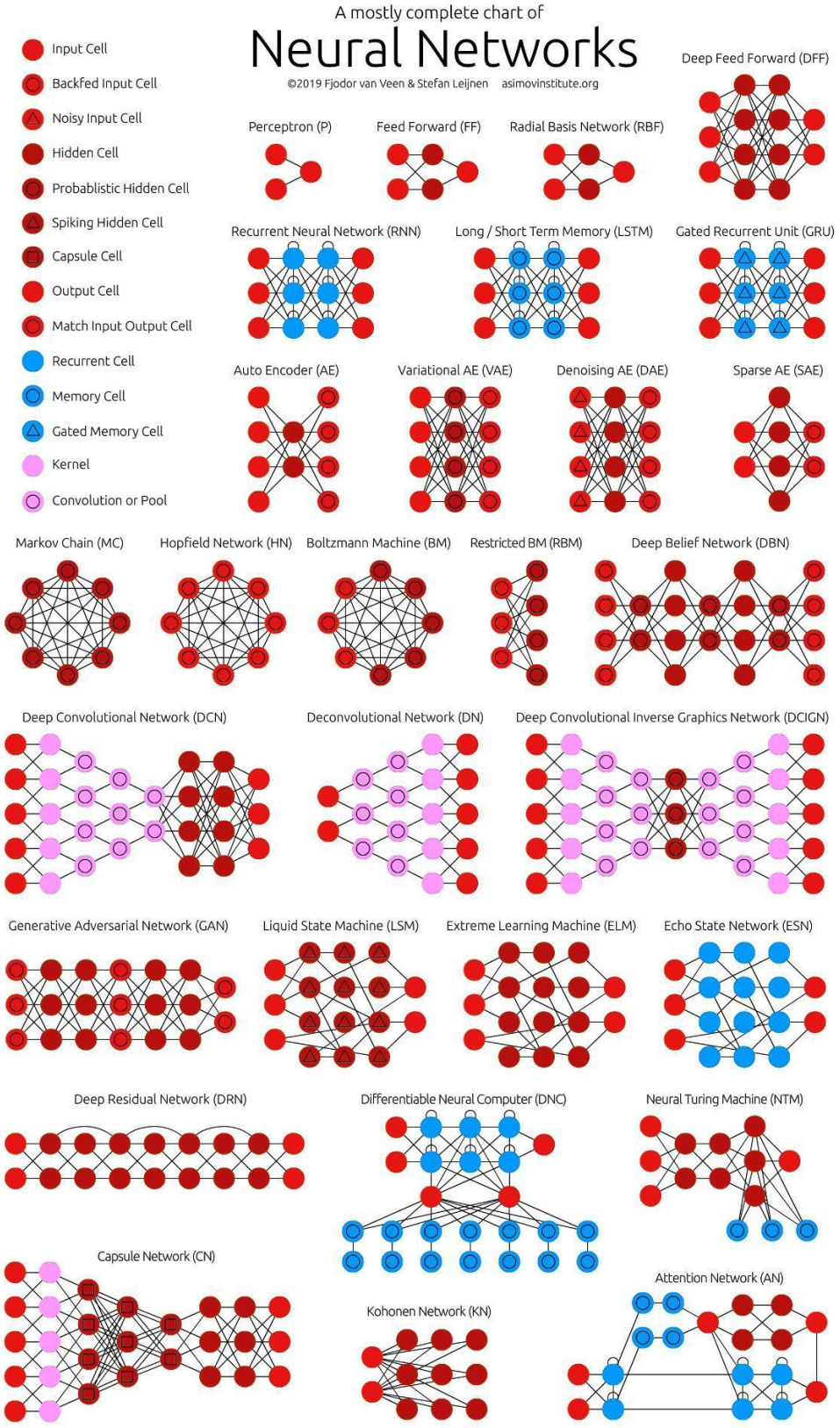

Рисунок 2.10 Класифікація архітектур нейронних мереж [28]

Крім відмінностей у типах зв'язку, архітектури нейронних мереж можуть відрізнятися кількістю прихованих шарів і числом нейронів. Незважаючи на універсальну теорему апроксимації, деякі функції дуже важко або майже неможливо точно апроксимувати лише одним прихованим шаром.

Для точної роботи мережі оптимальна кількість шарів та нейронів зазвичай підбирається експериментально. Якщо їх замало, мережа неспроможна правильно визначити залежність (underfitting), аналогією може бути апроксимація квадратичної залежності лінійної функції. Якщо шарів чи нейронів дуже багато, мережа буде перебудовуватися, тобто. вона буде «пам'ятати» випадки з навчальної вибірки, коли помилка була, але не зможе дати точні результати для навчальної вибірки або реальних значень. Крім того, така мережа повільніша і потребує більше часу на навчання.

Однією з перших досі популярних архітектур є багатошаровий персептрон (рис. 2.11), який є повнозв'язною мережею прямого поширення.

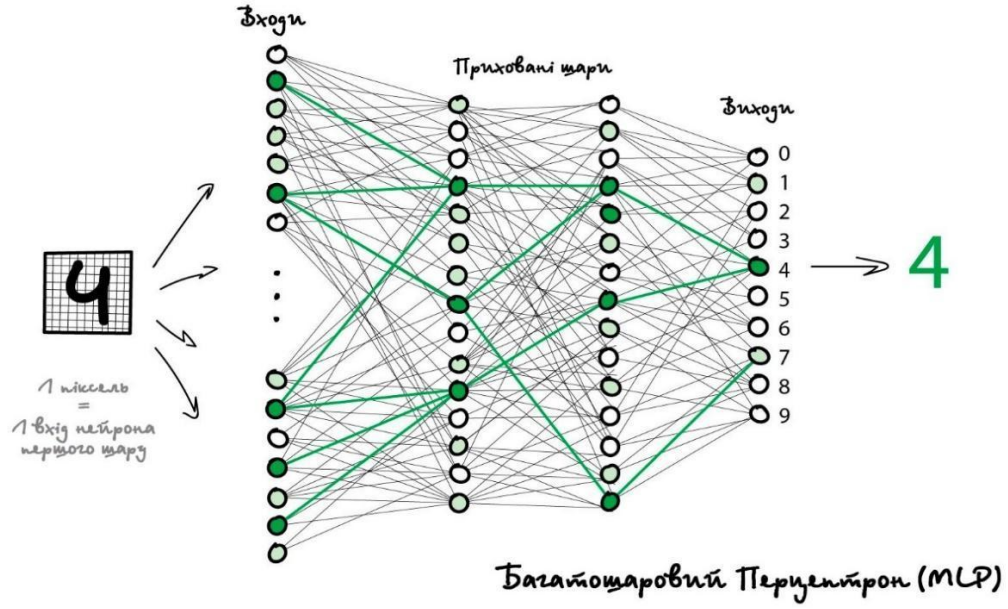

Рисунок 2.11 Приклад багатошарового перцептрону [23]

Нейронна мережа прямого поширення - це мережа, в якій сигнал йде тільки в одному напрямку - від вхідного шару, через приховані шари до вихідного шару, де результат виходить у вигляді скаляра або вектора.

Повнозв’язна нейронна мережа - це мережа, в якій кожен нейрон зв’язаний з усіма нейронами попереднього шару.

Переваги багатошарового персептрону:

- проста архітектура;

- універсальність – може вирішувати різні завдання.

Недоліки:

- для складніших завдань буде занадто багато параметрів, що робить його використання нераціональним;

- при великій кількості прихованих шарів відбувається «затухання» градієнта, що призводить до значного збільшення часу навчання.

З цього можна зробити висновок, що багатошаровий персептрон добре підходить для простих задач класифікації або регресії, а для складніших завдань, таких як розпізнавання зображень, слід вибрати нейронну мережу з іншою архітектурою. Така архітектура мережі підходить для вирішення нашого завдання.

Для того, щоб нейронна мережа функціонувала правильно, необхідно мати правильний баланс між усіма нейронами. Для отримання таких ваг використовуються алгоритми навчання. Залежно від архітектури мережі алгоритми навчання можна класифікувати так:

- навчання з учителем;

- навчання без вчителя;

- змішане навчання;

- навчання з підкріпленням.

Перед навчанням ваги мають бути ініціалізовані певними значеннями. Для цього існує кілька можливостей:

- використовувати ваги з попередньо навчених нейронних мереж;

- вибрати малі значення випадковим чином.

Найбільш відомим алгоритмом навчання є метод зворотного поширення помилки. Суть методу у тому, що результат порівнюється з реальним значенням у його отримання. Потім обчислюється помилка і у зворотному напрямку, отримується внесок кожного нейрона у мережу. Виходячи із цих вкладів, кожна вага коригується окремо.

Алгоритм методу:

1. Ініціалізувати ваги ($w_{ij} w_{ij}$), та $\Delta w_{ij} \Delta w_{ij} = 0$;
2. Повторювати доки помилка > є або досягнута певна кількість ітерацій:

Для усіх прикладів з навчальної вибірки:

1. подати вхідні значення та отримати вихідні значення з кожного нейрону ($a_{\downarrow} i)$ $a_{\downarrow} i)$;
2. для кожного вихідного нейрону розраховується помилка $\delta_k = f^{'}(a_k) * (y_k - a_k) \delta_k = f^{'}(a_k) * (y_k - a_k)$, де $f' f'$ – похідна від функції активації, $y_k y_k$ – реальне значення, $k \in k \in$ кількість вихідних нейронів;
3. для кожного наступного шару $i\ (i \in \text{кількість шарів} - 1)$ $i\ (i \in \text{кількість шарів} - 1)$: Для кожного нейрону $j\ (j \in \text{кількість нейронів у } i) j\ (j \in \text{кількість нейронів у } i)$ розраховується
4. $\delta_j = f^{'}(a_j) * \sum_{k \in \text{нейрони з } i+1} \delta_k w_{j,k} \delta_j = f^{'}(a_j) * \sum_{k \in \text{нейрони з } i+1} \delta_k w_{j,k}$.
5. для кожного $w_{ij} w_{ij}$:
6. $\Delta w_{ij} = \alpha * \delta_j * a_i \Delta w_{ij} = \alpha * \delta_j * a_i$; $w_{ij} := w_{ij} + \Delta w_{ij}$ $w_{ij} := w_{ij} + \Delta w_{ij}$,
7. де $\alpha\ \alpha$ – параметр швидкості навчання;
8. повернути значення $w_{ij} w_{ij}$.

2.1.4 Random forest

Random forest – ансамблевий метод машинного навчання, який полягає у використанні ансамблю дерев рішення. Крім того, метод поєднує у собі дві ідеї: метод беггінгу та метод випадкових підпросторів [32]. Алгоритм використовується для задач класифікації та регресії.

Дерева рішень мають два типи об'єктів - вузли та листя. Вузли містять правила, які використовуються для перевірки об'єктів та поділу групи об'єктів на підгрупи. Листя - це кінцеві вузли дерева, що містять підмножини, пов'язані з певними класами. Основна відмінність листя від вузлів полягає у відсутності огляду та подальшого розгалуження [33].

Як і інші моделі, дерева рішень будуються на навчальній вибірці. Коли дерево побудоване, формуються правила прийняття рішень, і для кожного такого правила створюється вузол. Для кожного вузла необхідно вибрати ознаку (атрибут), яка використовуватиметься для перевірки правила. Атрибути повинні бути обрані таким чином, щоб забезпечити найкращу ефективність. Найкраще розбиття є те, що дозволяє класифікувати якомога більше об'єктів і створити «чисті» підмножини [33].

Ансамблеві методи - це методи, які використовують кілька моделей одночасно для досягнення більшої точності, ніж кожна модель окремо [35]. Залежно від типу ансамблю, моделі можуть бути засновані на одних і тих самих або різних методах навчання. Моделі, які використовуються в ансамблі, повинні відрізнятися один від одного. З іншого боку, зазвичай вибираються «нестабільні» методи навчання, такі як датчики викидів чи аномалій.

Ефективність ансамблевих методів можна оцінити за допомогою теореми Кондорсе «Про присяжних» [36]. Якщо кожен присяжний має незалежну думку та ймовірність того, що присяжний прийме правильне рішення, більше 0,5, то ймовірність того, що присяжний прийме правильне рішення, зростає зі збільшенням числа присяжних та наближається до одиниці. Математично це можна записати так:

$$\mu = \sum_{i=m}^{N} C_N^i p^i (1-p)^{N-i} \mu = \sum_{i=m}^{N} C_N^i p^i (1-p)^{N-i},$$

де μ μ – ймовірність правильного рішення, N N – кількість присяжних, m m – мінімальна більшість членів журі ($m = \left\lfloor \frac{N}{2} \right\rfloor + 1$ $m = \left\lfloor \frac{N}{2} \right\rfloor + 1$), p p – ймовірність правильного рішення присяжного. Отже, якщо $p > 0.5$ $p > 0.5$, то $\mu > p$ $\mu > p$. А якщо $N \to \infty$ $N \to \infty$, то $\mu \to \infty$ $\mu \to \infty$.

Беггінг (Bagging, від Bootstrap aggregation) – один з видів ансамблів, який базується на статистичному методі бутстрепу [37]. Він полягає у наступному. Нехай існує вибірка X X розміром N N. З вибірки випадково вибирається N N об'єктів з поверненням й створюється підвибірка $X_1 X_1$. Таким чином, на кожній спробі у кожного об'єкта ймовірність щоразу бути вибраним складає. Об'єкти у підвибірці можуть повторюватися. Повторюємо процедуру разів M M й генеруємо підвибірки $X_1, \ldots, X_M X_1, \ldots, X_M$.

Варто зазначити, що при генеруванні підвибірки на основі методу бутстрепу приблизно 37% об'єктів з початкової вибірки не потрапляють у неї й такі об'єкти називаються Out-of-bag. Математично це можна довести таким чином. Нехай у вибірці N N об'єктів. На кожному кроці кожен об'єкт потрапляє у підвибірку з ймовірністю. Отже, ймовірність того що об'єкт не попаде у підвибірку складає $\left(1 - \frac{1}{N}\right)^N \left(1 - \frac{1}{N}\right)^N$. При $N \to \infty$ $N \to \infty$ отримуємо ймовірність . А ймовірність кожного об'єкту попасти у підвибірку складає $1 - \frac{1}{e} \approx 63\%$ $1 - \frac{1}{e} \approx 63\%$.

Суть беггінгу виглядає наступним чином (рис. 2.12). На основі вибірки X X генеруються підвибірки $X_1, \ldots, X_M X_1, \ldots, X_M$. На кожній підвибірці тренується модель $a_i(x) a_i(x)$. Кінцева модель буде усереднювати відповіді кожної моделі:

$a(x) = \frac{1}{M}\sum_{i=1}^{M} a_i(x) a(x) = \frac{1}{M}\sum_{i=1}^{M} a_i(x)$, або у випадку класифікації проводити голосування.

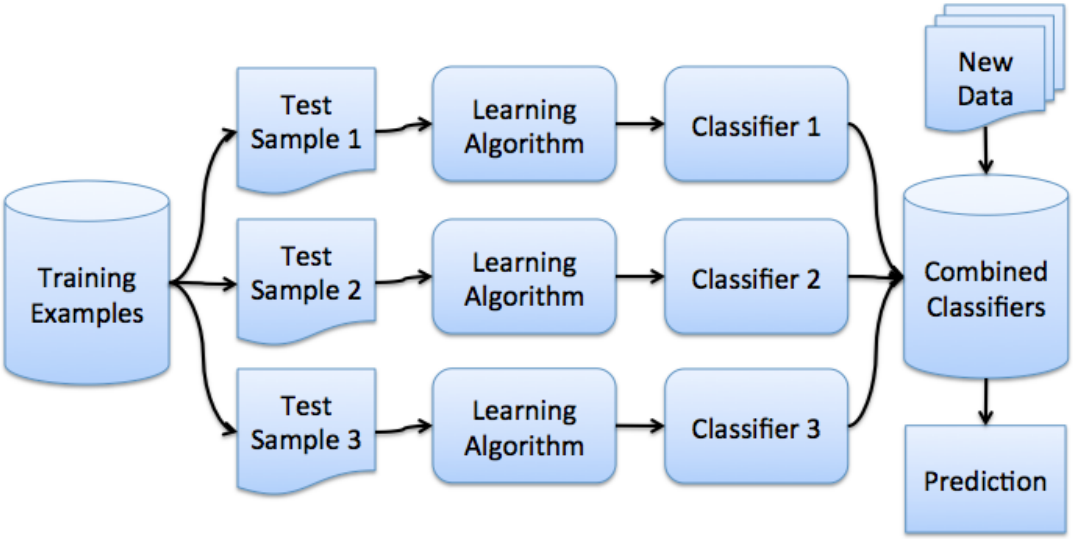

Рисунок 2.12 Приклад ансамблевого методу на основі беггінгу [38]

Ефективність беггінгу забезпечується за рахунок того, що моделі, навчені на різних підвиборках, абсолютно різні та їх помилки взаємно компенсуються. Крім того, викиди даних можуть не потрапляти до деяких підвиборок. Беггінг ефективний на невеликих вибірках, коли втрата навіть невеликої частини об'єктів призводить до появи різних моделей. У разі великих вибірок можна генерувати підвибірки меншого розміру.

Крім беггінгу, випадковий ліс застосовує метод випадкових підпросторів [39]. У цьому методі моделі навчаються на основі підмножини ознак, які вибираються випадковим чином. Цей метод зменшує кореляцію між деревами та знижує ймовірність перенавчання. Алгоритм побудови ансамблів, заснований на методі випадкових підпросторів, має такий вигляд:

1. нехай є N об’єктів у вибірці D ознак, й M моделей у ансамблі;
2. для кожної моделі $a_i(x)$ обирається d ознак , при цьому $d < D$;

3. для кожної моделі $a_i(x) a_i(x)$ створюється підвибірка, обираючи випадковим чином $d\ d$ ознак з $D\ D$ й проводиться навчання;
4. на основі отриманих моделей утворюється ансамбль, результат у якому отримується шляхом усереднення відповідей: $a(x) = \frac{1}{M}\sum_{i=1}^{M} a_i(x) a(x) = \frac{1}{M}\sum_{i=1}^{M} a_i(x)$, або шляхом голосування у випадку класифікації.

У випадковому лісі для задач регресії рекомендується обирати $d = \frac{D}{3}$ $d = \frac{D}{3}$, а для задач класифікації $d = \sqrt{D} d = \sqrt{D}$. Проте у різних задач оптимальне значення $d\ d$ може різнитися й підбирається експериментальним методом [40].

Таким чином, об’єднавши беггінг на деревах рішень та метод випадкових підпросторів отримуємо випадковий ліс. Алгоритм випадкового лісу такий:

1) Для кожного $n = 1, \ldots, N\ n = 1, \ldots, N$, де $N\ N$ – необхідна кількість дерев у ансамблі:

1. згенерувати підвибірку $X_n X_n$ за допомогою бутсрепу з вибірки $X\ X$;
2. побудувати дерево рішень $a_i(x) a_i(x)$ на основі підвибірки $X_n X_n$:

- по заданому критерію обирається краща ознака, по якій проводиться розбиття у дереві до вичерпання підвибірки;
- при кожному розбитті обирається $d\ d$ випадкових ознак серед $D\ D$, й оптимальне розділення вибірки шукається лише серед них;
- дерево будується поки не досягається певна висота, або по досягненню певної кількості об’єктів у листках.

2) З отриманих дерев рішень сформувати ансамбль, результат якого буде формуватися за виразом $a(x) = \frac{1}{N}\sum_{i=1}^{N} a_i(x) a(x) = \frac{1}{N}\sum_{i=1}^{N} a_i(x)$.

2.2 Аналіз основних видів вхідних даних та їх вибір

2.2.1 Вибір вхідних даних

Як дані для сільського господарства використовуються кількісні (числові) дані, такі як добрива, опади, температура тощо. Останні десятиліття дані космічних спостережень чи дистанційного зондування Землі, наприклад, з допомогою безпілотників, разом із наземної інформацією стають дедалі більше корисними моніторингу врожайності, зокрема сільськогосподарських культур. Спостереження з космосу за допомогою супутників з використанням чутливих датчиків за параметрами та властивостями земної поверхні ведуться з кінця минулого століття. Отримані дані відкривають нові можливості для моніторингу та контролю за врожайністю сільськогосподарських культур. На додаток до традиційних аерофотознімків використовуються складніші дані, такі як NDVI (Normalized Difference Vegetation Index), VHI (Vegetation Health Index) тощо.

При створенні моделей прогнозування такі складні показники дають значну перевагу, тому що за рахунок використання невеликої кількості показників розмірність вхідних даних невелика, що спрощує моделі та позитивно позначається на їхній точності при малих обсягах вибірки. Тому було вирішено використовувати NDVI і VHI як характеристики разом з кількістю внесених добрив.

2.2.2 Індекс NDVI

Нормалізований диференційний вегетаційний індекс (NDVI) є показником кількості фотосинтетичної активної біомаси [41].

NDVI широко використовується у всьому світі для моніторингу засухи, моніторингу та прогнозування сільськогосподарського виробництва, а також для прогнозування пожежонебезпечних зон та карт пустель. NDVI найкраще підходить для глобального моніторингу рослинності, оскільки він компенсує зміни умов освітленості, нахилу поверхні, експозиції та інших зовнішніх факторів [42].

NDVI розраховується за такою формулою:

$$NDVI = \frac{NIR - RED}{NIR + RED} NDVI = \frac{NIR - RED}{NIR + RED},$$

де NIR – кількість відбитого інфрачервоного випромінювання, RED – кількість відбитого червоного випромінювання.

Для відображення значень індексів використовується безперервний градієнт або дискретна шкала. Значення індексу NDVI коливається від -1 до 1. Значення індексу для рослинності зазвичай коливається від 0,2 до 0,95 (рис. 2.13). Чим краще розвинена рослинність протягом вегетації, тим вище значення НДВІ.

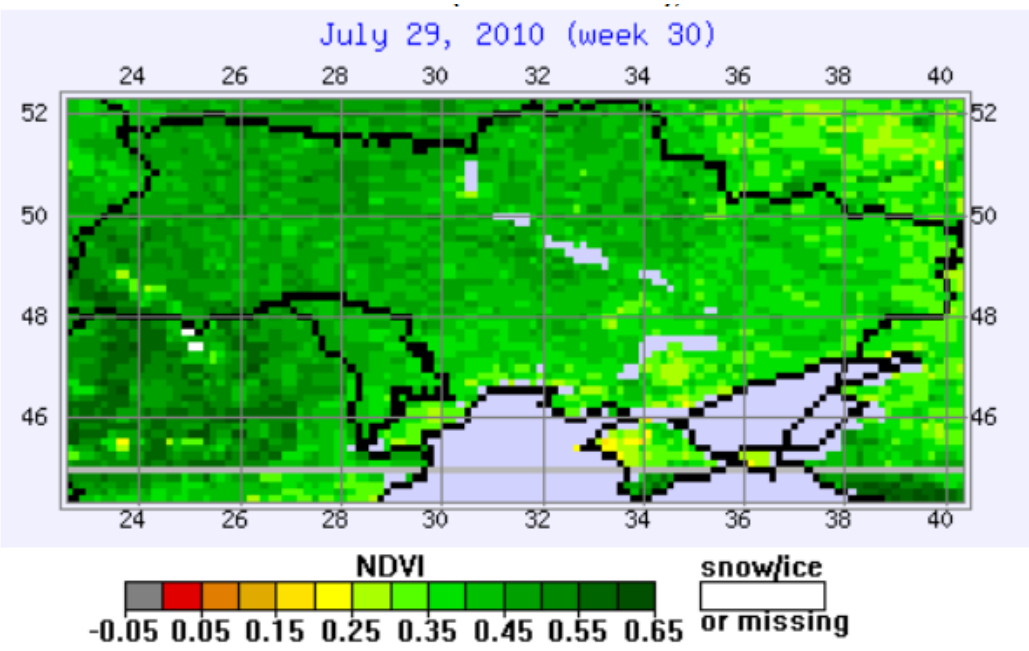

Рисунок 2.13 Приклад графічного відображення NDVI
з дискретною шкалою [43]

Таким чином, NDVI – це індекс, за яким можна судити про розвиток зеленої маси рослин протягом вегетаційного періоду. Оскільки відображувальна здатність невегетативних об'єктів не залежить від пори року, їх індекс NDVI має фіксоване значення, яке є нижчим порівняно з рослинами (рис. 2.14).

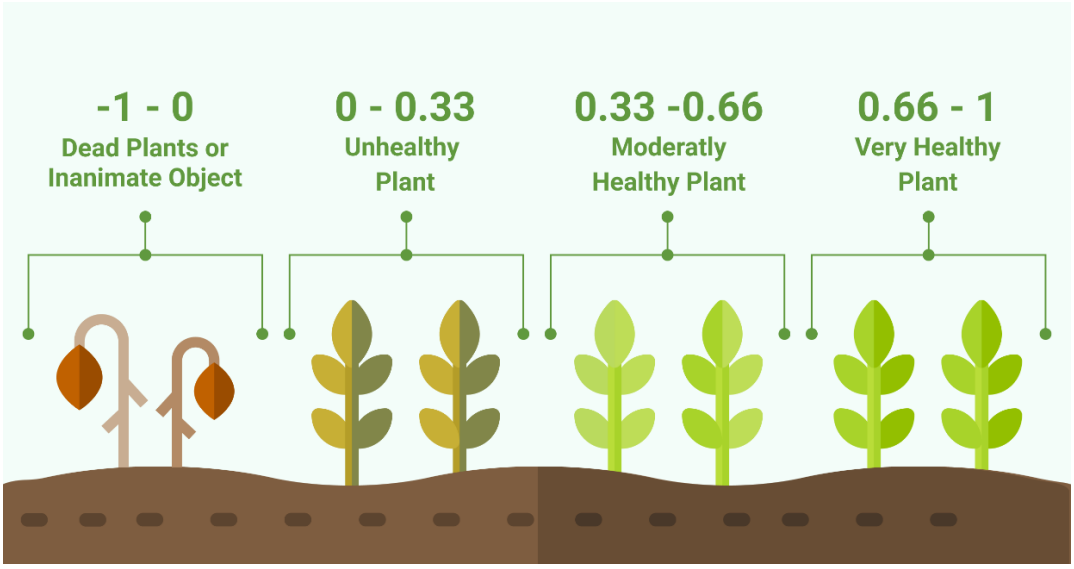

Рисунок 2.14 Залежність NDVI від стану здоров’я рослин [44]

Протягом вегетаційного періоду індекс NDVI зростає, досягає піку приблизно за 0,80-0,85 (у зернових це момент формування колосу), потім починає знижуватися (рис. 2.15). Зниження індексу наприкінці вегетації відбиває зрілість рослин. Наприклад, оптимальний порядок збирання полів можна визначити для кількох зернових полів за індексом NDVI - чим нижчий індекс, тим сухіше зерно.

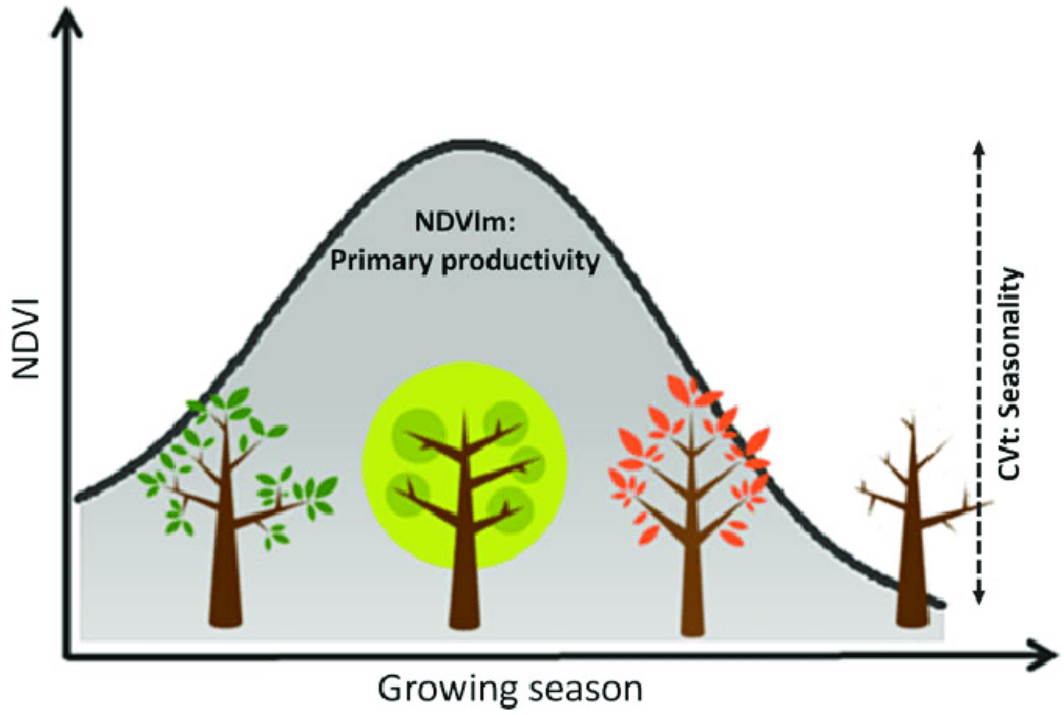

Рисунок 2.15 Вплив фази розвитку рослин на індекс NDVI [45]

Таким чином, індекс NDVI добре підходить для оцінки розвитку сільськогосподарських культур і може використовуватися в системах прогнозування врожайності.

2.2.3 Індекс VHI

Вегетаційний індекс здоров'я (VHI) є показником, який характеризує здоров'я рослинності, і свідчить про те, що стресові умови рослин пов'язані з нижчим за нормальний рівень NDVI і вищою за нормальну температуру [46, 47].

Розраховується VHI за формулою:

$$VHI = \alpha * VCI + (1 - \alpha) * TCI$$

Зазвичай використовується як індекс посухи на основі даних дистанційного зондування, VHI розраховується як зважена сума двох компонентів: індексу стану рослинності (VCI) та індексу теплового стану

(TCI). Використання комбінації індексів TCI та VCI забезпечує покращене представлення рівнів засухи [48].

VCI характеризує рівень вологості і зазвичай ґрунтується на даних видимого та ближнього інфрачервоного електромагнітного спектру. Цей показник є високоточним як оцінка засухи і заснований на його впливі на тривалість та інтенсивність вегетаційного періоду [48]. VCI можна використовувати разом з іншими індикаторами для прогнозування стану рослинності. Індекс розраховується за такою формулою:

$$VCI = \frac{NDVI' - NDVI_{min}}{NDVI_{max} - NDVI_{\min}\square} VCI = \frac{NDVI' - NDVI_{min}}{NDVI_{max} - NDVI_{\min}\square},$$

де $NDVI'NDVI'$ – середнє значення NDVI, $NDVI_{\min}\square NDVI_{\min}\square$ – найменше значення індексу, $NDVI_{\max}\square NDVI_{\max}\square$ – найбільше значення за певний період спостереження.

Значення VCI знаходяться в діапазоні між 0 та 1. Низькі значення вказують на стресові умови рослинності, а високі – на оптимальний стан.

TCI характеризує температурний рівень на основі даних з інфрачервоного спектру. Даний індекс забезпечую кращу оцінку стресу рослинності на основі температурних показників [48].

Розраховується TCI за формулою:

$$TCI = \frac{BT_{max} - BT'}{BT_{max} - BT_{\min}\square} TCI = \frac{BT_{max} - BT'}{BT_{max} - BT_{\min}\square},$$

де $BT'BT'$ – середнє композитне значення температури, $BT_{\max}\square BT_{\max}\square$ – найбільше значення температури, $BT_{\min}\square BT_{\min}\square$ – найменше значення температури за певний період спостереження.

Оптимальні ваги для VCI та TCI під час розрахунку VHI зазвичай невідомі, тому вагові коефіцієнти для кожного індексу приймаються рівними 0,5, отже α = 0,5. Останні дослідження показали, що можна підвищити точність ваг, порівнюючи отримані VHI і несупутникові показники засухи, такі як SPEI [49].

Аналіз даних показує кореляцію між TCI та VCI (LST-NDVI). Сильна кореляція дозволяє використовувати ці індекси як основу розрахунку VHI. Однак при слабкій кореляції використання TCI та VCI не підходить, тому VHI не може бути розрахований, тому для деяких районів цей індекс не може бути використаний як оцінка засухи [50].

Для відображення значення індексу використається безперервний градієнт або дискретна шкала. Значення індексу VHI варіює від 0 до 100 (рис. 2.16). Чим вище значення індексу, тим здоровіша рослинність.

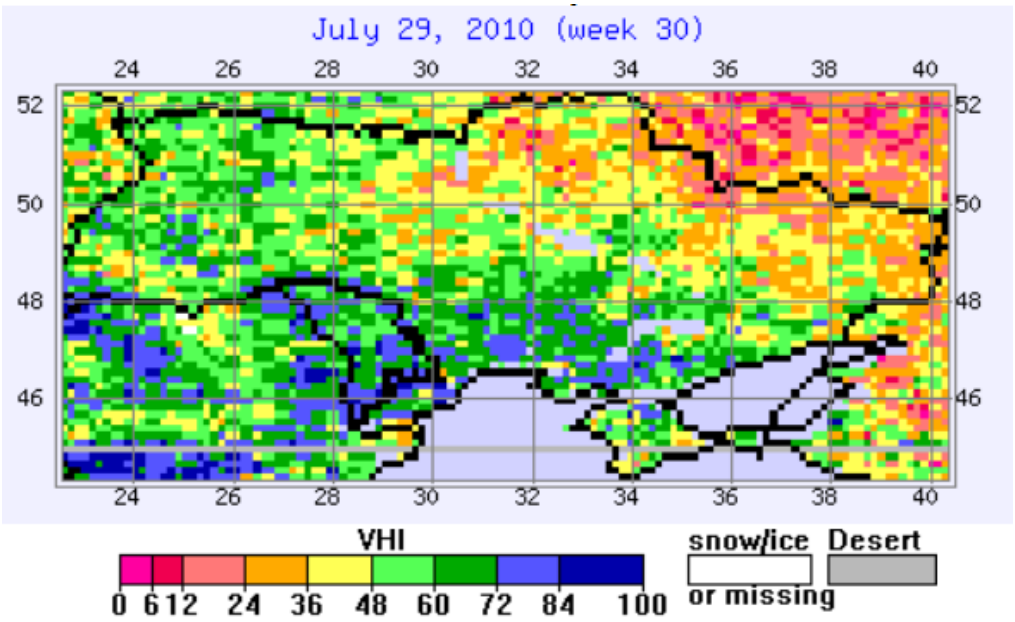

Рисунок 2.16 Приклад графічного відображення VHI
з дискретною шкалою [43]

Таким чином, VHI можна використовувати як вхідні дані для прогнозування врожайності сільськогосподарських культур.

РОЗДІЛ 3 СТВОРЕННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ, ОПИС ТА РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТУ

3.1 Архітектура моделей прогнозування врожайності сільськогосподарських культур

3.1.1 Попередній аналіз даних

Перш ніж почати будувати моделі для прогнозування врожайності, корисно проаналізувати дані, використані при створенні вибірок. Результати аналізу можуть дати підказки та допомогти з вибором необхідної архітектури.

Як уже зазначалося, NDVI, VHI та кількість внесених добрив (кг/га) використовуються як вхідні дані (незалежні змінні). Як вихідні дані (залежна змінна) використовується врожайність конкретної сільськогосподарської культури (ц/га). В якості культур відібрали зернові, ріпак та кормові буряки. Для України дані були зібрані за період 1992-2021 рр.

Через невеликий розмір вхідних даних першим кроком аналізу є їх візуалізація (рис. 3.1-3.3). Для візуалізації використовуються матриці графіків. На діагоналі матриці розташовані гістограми розподілу ознак, а інші елементи матриці містять діаграми розсіювання відповідних пар ознак. Візуалізація використовує бібліотеку Seaborn мови Python на основі бібліотеки Matplotlib і бібліотеку Pandas для взаємодії з даними.

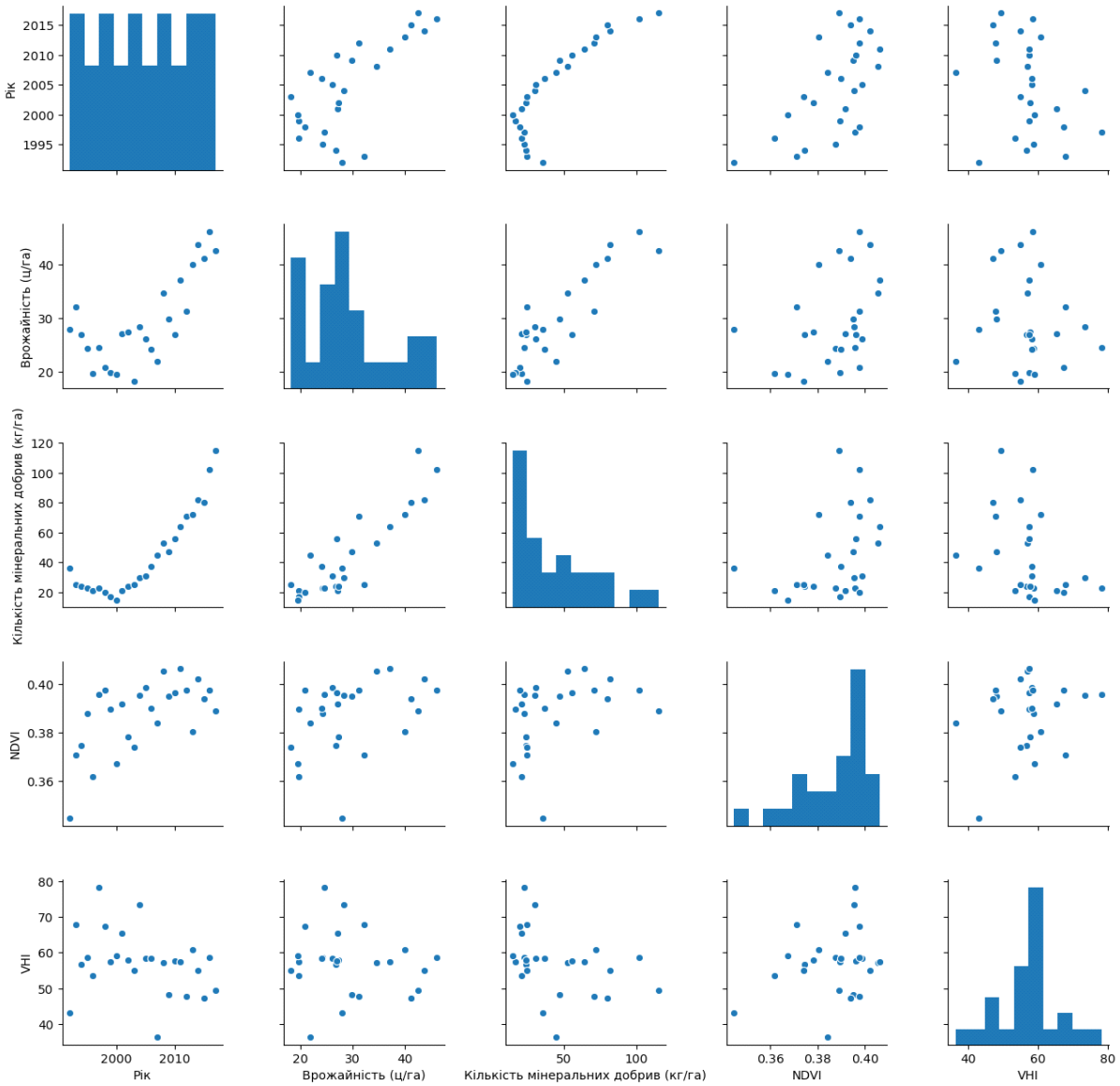

Рисунок 3.1 Матриця графіків для зернових культур

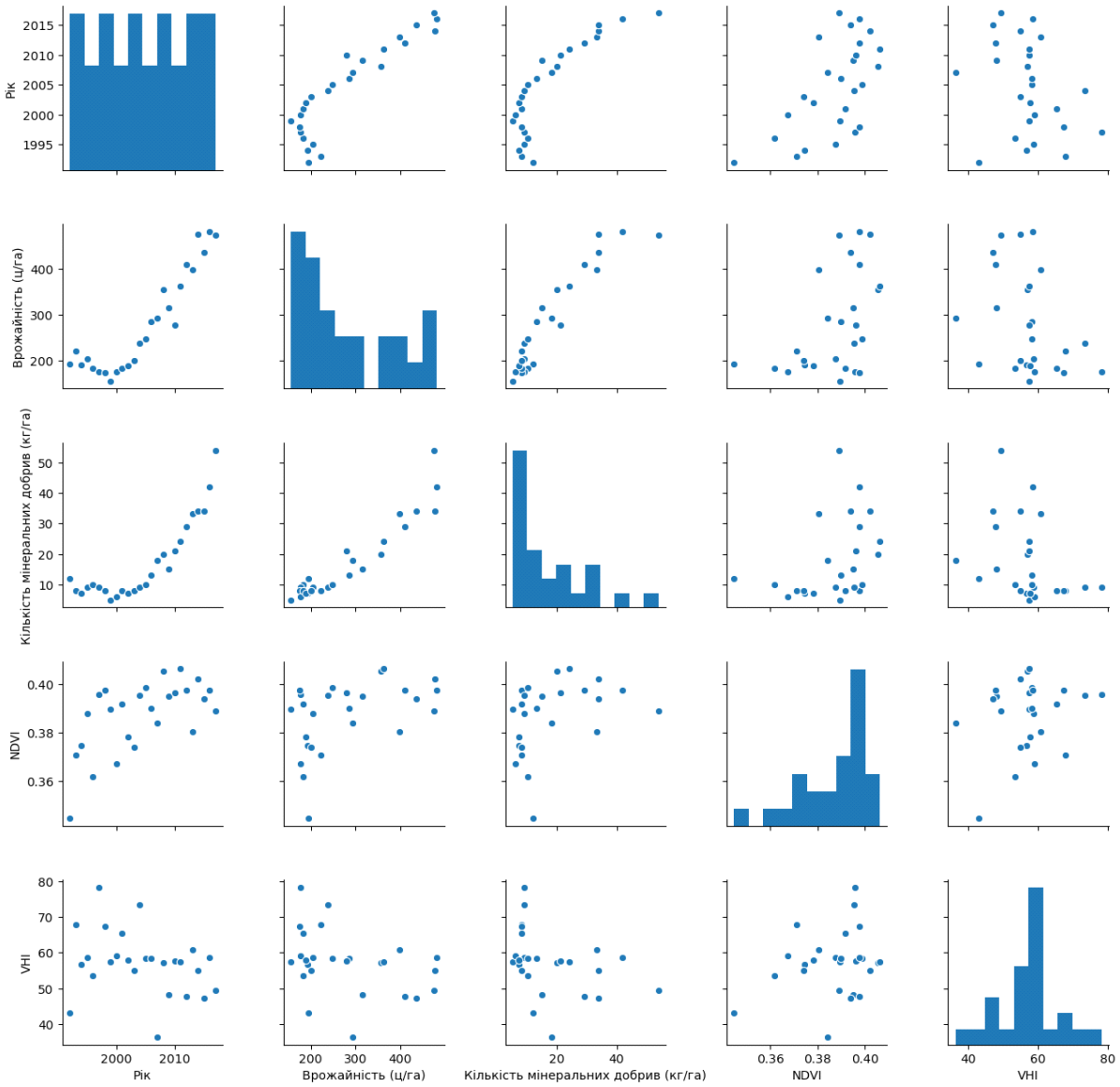

Рисунок 3.2 Матриця графіків для кормового буряку

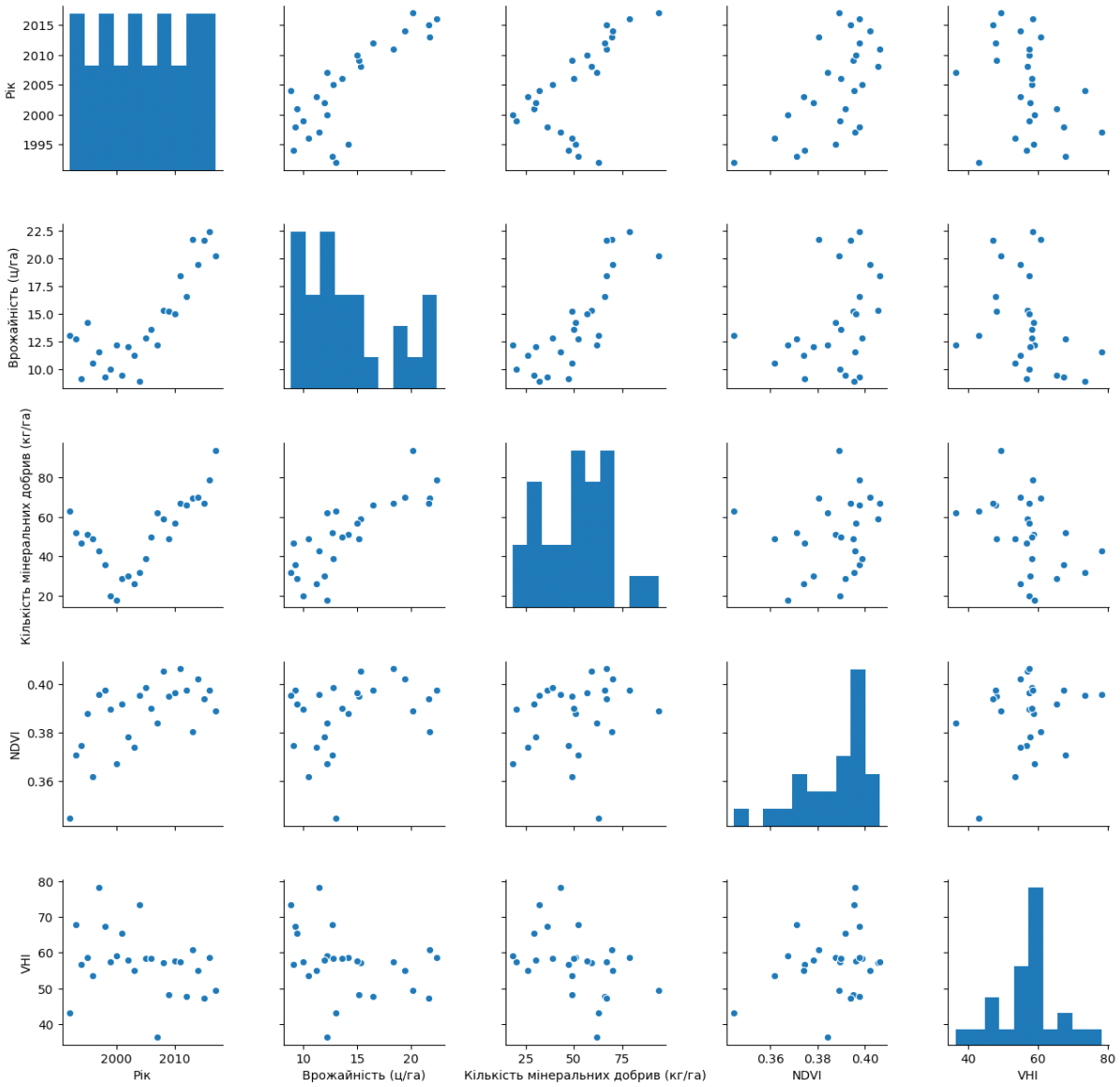

Рисунок 3.3 Матриця графіків для ріпаку

При візуальному аналізі отриманих результатів можна відмітити досить чітку лінійну залежність між урожайністю та кількістю добрив, що вносяться під ці культури. Щодо інших важливих показників, на основі візуального аналізу неможливо зробити однозначних висновків про залежності.

Цікаво, що між парами характеристик «рік» — «кількість мінеральних добрив» і «рік» — «врожайність» є дуже чітка нелінійна залежність. Подібний зв'язок вже виявлено іноземними дослідниками, а саме позитивна лінійна динаміка врожайності озимої пшениці за останні десятиліття, що пов'язана з удосконаленням агротехніки [51]. Можна припустити, що

нелінійна залежність пов’язана з економічними проблемами України на початку її незалежності.

Для більш детального аналізу використовуємо коефіцієнт кореляції Пірсона, який показує ступінь лінійної залежності між залежною та незалежною змінними. Для обчислення кореляції використовувалася бібліотека Numpy. Отримані результати наведено в табл. 3.1.

Таблиця 3.1 Результати кореляційного аналізу

	Кількість мінеральних добрив	NDVI	VHI
Зернові культури	0,75	0,34	-0,10
Кормовий буряк	0,71	0,31	-0,31
Ріпак	0,80	0,43	-0,35

Кореляційний аналіз підтвердив сильну лінійну залежність між врожайністю та внесеними добривами. Інші показники мають слабку кореляцію, а VHI навіть негативний. Однак слабка кореляція може означати, що існує залежність нелінійного типу.

3.1.2 Вибір архітектури

Для вибору оптимальної архітектури тестуються різні її варіанти. Для подальших експериментів вибирається найкращий варіант архітектури.

Для тестування моделей було обрано метод перехресної перевірки. Суть процедури полягає в поділі зразка на K K блоків однакового розміру. Один з блоків використовується як тестова вибірка, а інші $K-1$ $K-1$ утворять навчальну вибірку. На утвореній навчальній вибірці відбудеться навчання моделі, а для оцінки точності використовується тестова вибірка. Процес повторюється K K разів, причому кожен блок буде

використовуватися як тестова вибірка лише один раз. У результаті буде отримано $K\ K$ оцінок, а результуюча оцінка розраховується шляхом усереднення. Для кожної перевірки буде використовуватися $K = 5\ K = 5$.

У якості оцінки буде використано середньоквадратичну помилку, яка розраховується за формулою $MSE = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$

$MSE = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$, де $n\ n$ – кількість випадків, $y^{(i)}y^{(i)}$– реальне значення у *i*-му випадку, $h_\theta\left(x^{(i)}\right)h_\theta\left(x^{(i)}\right)$ – прогнозоване значення для *i*-го випадку. Таким чином, чим нижча оцінка, тим кращий результат.

Використовуваним програмним забезпеченням є бібліотека Scikit-learn, яка містить усі описані вище методи машинного навчання та необхідні інструменти для тестування. Крім того, для роботи з даними використовується бібліотека Pandas.

Почнемо з поліноміальної регресії. Аналіз даних показав, що не всі незалежні та залежні змінні мають чітку лінійну залежність, тому проста лінійна регресія може бути не найкращим варіантом. Для перевірки використовуються поліноміальні регресії різних ступенів, починаючи з першого (лінійний варіант). Результати наведені в табл. 3.2.

Таблиця 3.2 Результат тестування поліноміальної регресії різних степенів

Степінь поліному	MSE		
	Зернові культури	Кормовий буряк	Ріпак
1	16,1	1630	15,1
2	$1{,}4 * 10^5$	$4{,}4 * 10^9$	$9{,}1 * 10^4$
3	$1{,}2* 10^{13}$	$7{,}6 * 10^{10}$	$2{,}2 * 10^6$

На основі отриманих результатів можна зробити висновок, що для моделювання врожайності кожної культури оптимальними є лінійні моделі (поліноми першого ступеня). Зі збільшенням степеня полінома похибка збільшується на кілька порядків, що свідчить про перенавчання моделі.

Наступним методом перевірки буде нейронна мережа. При виборі архітектури необхідно вдало вибрати кількість нейронів у прихованому шарі, щоб їх було достатньо, щоб знайти залежність, але ненабагато, щоб уникнути перенавчання. ReLU використовується як функція активації, оскільки попередні тести показали, що при використанні сигмоїдної функції мережа навчається повільніше й дає більшу помилку. Для зернових культур та соняшнику максимальна кількість ітерацій буде 10000, а для кормового буряку – 25000. Параметр швидкості навчання мережі – 0,001. Результати тестування занесені до табл. 3.3.

Таблиця 3.3 Результат тестування нейронних мереж з різною кількістю нейронів у прихованому шару

Кількість нейронів у прихованому шарі	MSE		
	Зернові культури	Кормовий буряк	Ріпак
3	72,5	7123	18,2
5	16,2	2799	9,1
7	20,2	2404	9,5
9	20,1	1946	10,7
12	21,8	2836	10,2

Аналізуючи результати, можна побачити загальну тенденцію: при малій кількості нейронів похибка висока. Зі збільшенням числа похибка зменшується до певного рівня, а потім знову починає збільшуватися. Таким чином, оптимальну кількість нейронів у точці мінімальної похибки

визначити легко. Таким чином, оптимальні архітектури нейронних мереж для кожної культури є такими:

- Зернові культури – 3/5/1;
- Кормовий буряк – 3/9/1;
- Ріпак – 3/5/1.

Останній метод тестування — Random Forest. Перевага випадкового лісу полягає в тому, що немає перенавчання, якщо є занадто багато дерев рішень. Однак, якщо дерев буде замало, точність моделі буде недостатньою, а якщо їх кількість збільшується, модель стане складнішою, потребуватиме більше пам’яті та працювати повільніше. Тому необхідно оптимально підібрати кількість дерев у «лісі». Результати тесту наведені в табл. 3.4.

Таблиця 3.4 Результати тестування випадкового лісу з різною кількістю дерев рішень

Кількість дерев	MSE		
	Зернові культури	Кормовий буряк	Ріпак
10	59,4	5253	10,5
25	57,5	5177	9,1
50	56,8	4858	9,7
100	56,5	4809	9,8

Результати підтвердили, що зі збільшенням кількості дерев рішень точність моделі зростає, але швидкість зростання точності поступово зменшується. Різниця в результатах від 50 до 100 одиниць на одну культуру досить мала, що свідчить про недоцільність подальшого збільшення кількості. Тому для кожної моделі на базі випадкового лісу використовується 100 дерев рішень.

3.2 Опис програмного додатку

3.2.1 Розробка програмного додатку

Щоб розробити програму, спочатку потрібно побудувати й навчити моделі на основі різних методів машинного навчання. Використовуйте такі методи:

- лінійна регресія;
- нейронна мережа;
- випадковий ліс.

Для кожного виду сільськогосподарської культури необхідно створити окрему модель. Оскільки в роботі використовуються 3 види рослин, необхідно зробити 9 різних моделей.

Параметри та алгоритми навчання лінійної регресії та моделі випадкового лісу однакові для всіх культур. У випадковому лісі використовується 100 одиниць.

Моделі на основі нейронних мереж мають відмінності:

- Зернові культури: архітектура 3/5/1, максимальна кількість епох – 10000, параметр швидкості навчання – 0,001, функція активації – ReLU;
- Кормовий буряк: архітектура 3/9/1, максимальна кількість епох – 100000, параметр швидкості навчання – 0,001, функція активації – ReLU;
- Ріпак: архітектура 3/5/1, максимальна кількість епох – 10000, параметр швидкості навчання – 0,001, функція активації – ReLU;

Оптимальність обраних параметрів описано в підрозділі 3.1.2.

В якості вхідних даних (незалежні змінні) використовуються: кількість внесених мінеральних добрив (кг/га), індекс NDVI, індекс VHI. В якості вихідних даних (залежна змінна) використовується врожайність культур (ц/га). Дані отримані зі звітів Державної служби статистики України [52] та супутникових даних NOAA STAR [43] (дані для України за 1992-2019 рр.).

Бібліотека Scikit-learn використовується для побудови моделей, оскільки містить реалізації всіх необхідних методів, а бібліотека Pandas

використовувалася для роботи з даними. Приклад коду для завантаження даних, побудови та вивчення моделі показано на рис. 3.4.

```
>>> import pandas as pd
>>> from sklearn.linear_model import LinearRegression
>>> data = pd.read_excel('Зерно.xlsx')
>>> X = data[['Добрива','NDVI','VHI']].to_numpy()
>>> y = data['Врожайність '].to_numpy()
>>> model = LinearRegression()
>>> model.fit(X, y)
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

Рисунок 3.4 Приклад створення моделі на основі лінійної регресії

Після створення та навчання моделей їх необхідно зберегти для подальшого використання. Для цього використовується вбудована бібліотека Pickle, яка дозволяє зберігати об’єкти Python у файлах. Приклад коду для зберігання моделі наведений на рис. 3.5.

```
>>> import pickle
>>> pickle.dump(model, open('model.sav', 'wb'))
```

Рисунок 3.5 Приклад коду для зберігання моделі

Після збереження моделей у вигляді файлів, їх було використано у розробленому програмному додатку. Для зворотного перетворення файлу у об’єкт використовується та сама бібліотека Pickle. Приклад коду для завантаження файлу та використання моделі для отримання результату наведено на рис. 3.6.

```
>>> model = pickle.load(open('model.sav', 'rb'))
>>> model.predict([[100, 0.38, 56]])
array([45.67322244])
```

Рисунок 3.6 Приклад коду завантаження та використання моделі

Для зручності використання програмного додатку використовується графічний інтерфейс, який реалізований у бібліотеці PySimpleGUI. Повний код програмного додатку наведений у додатку А.

Повний перелік програмного забезпечення та бібліотек, необхідний для роботи створеної програми:

- ОС: Windows 10
- Python 3.8
- Scikit-learn 0.22.2
- PySimpleGUI 4.18.2
- Pickle

3.2.2 Путівник з використання програмного додатку

Створене програмне забезпечення має інтуїтивно зрозумілий графічний інтерфейс (рис. 3.7), проте має певні особливості, які необхідно описати.

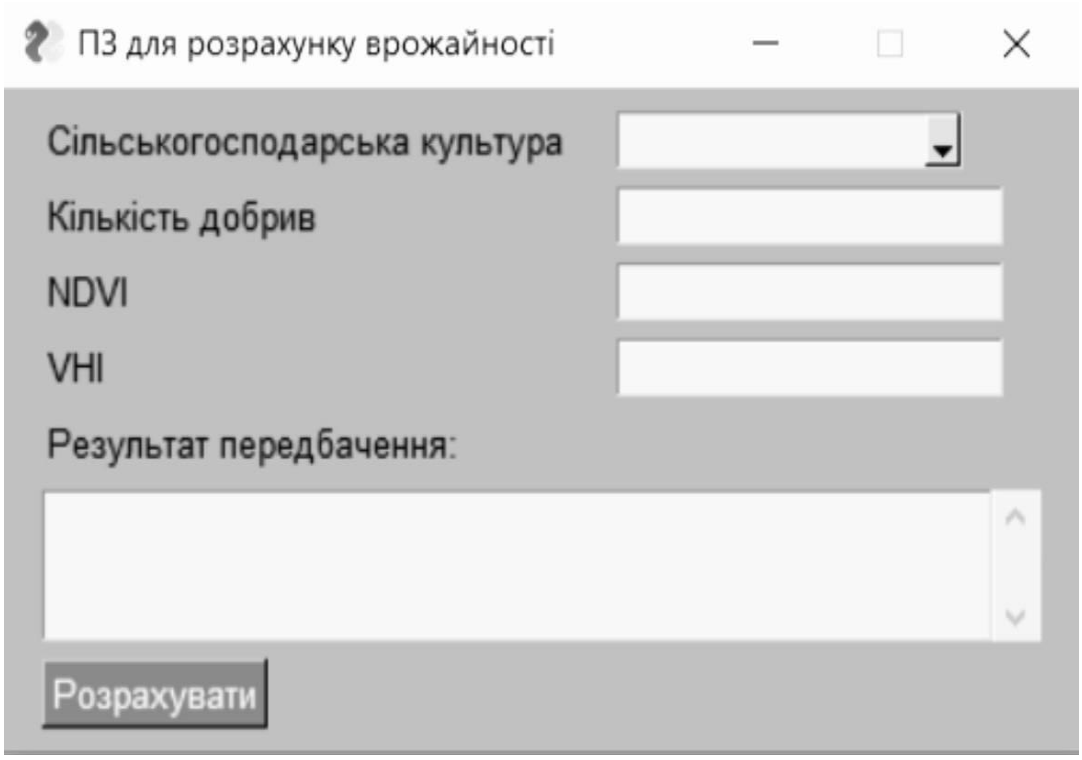

Рисунок 3.7 Інтерфейс ПЗ при запуску

Для початку роботи, слід вибрати необхідну для прогнозування сільськогосподарську культуру з випадаючого меню (рис. 3.8).

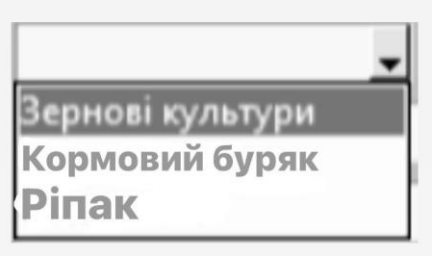

Рисунок 3.8 Випадаюче меню з переліком культур

Далі необхідно ввести у поле кількість внесених мінеральних добрив, одиниці вимірювання – кг/га. Потім, у наступне поле, показник NDVI, значення якого знаходяться в інтервалі між -1 та 1. Останнім є показник VHI, значення якого знаходяться в інтервалі між 0 та 100. Для зручності, перед кожним полем для заповнення міститься текст з відповідною назвою. Після внесення всіх необхідних даних необхідно натиснути на кнопку з написом «Розрахувати». При натисканні кнопки, результат прогнозування кожної моделі виводиться у вікні під текстом «Результат передбачення:» (рис. 3.9).

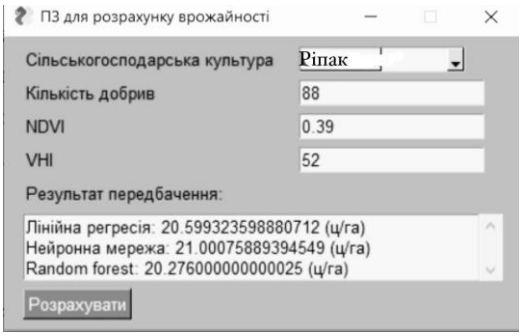

Рисунок 3.9 Приклад роботи ПЗ

При необхідності введення дробових значень потрібно використовувати десяткові дроби, а розділяти цілу й дробову частину крапкою. Якщо дані введено некоректно, з’явиться вікно з відповідним попередженням і ніяких розрахунків не відбудеться (рис. 3.10).

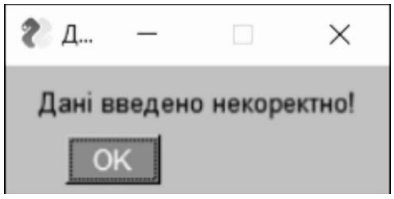

Рисунок 3.10 Вікно з попередженням про неправильність введених даних

А якщо не всі поля заповнені, то з’явиться вікно з відповідним попередженням та результат прогнозування не буде отримано (рис. 3.11).

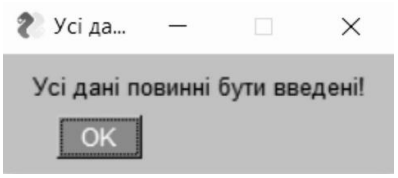

Рисунок 3.11 Вікно з попередженням про неповноту даних

3.3 Опис експерименту

Для проведення експерименту буде використовуватися розроблений програмний додаток.

Мета експерименту полягає у визначенні методу машинного навчання, моделі якого дадуть найточніше передбачення для різних сільськогосподарських культур.

Для оцінки якості моделі буде здійснено передбачення врожайності на 2018 рік, дані за який не використовувалися під час навчання, та розрахована точність прогнозу. Після проведення експериментів, отримані результати аналізуються та на його основі робляться висновки.

3.4 Результати експерименту

Розпочнемо з передбачення врожайності зернових культур. Реальний показник врожайності в 2018 році склав 47,4 ц/га.

Результати прогнозування врожайності зернових культур наведені на рис. 3.12.

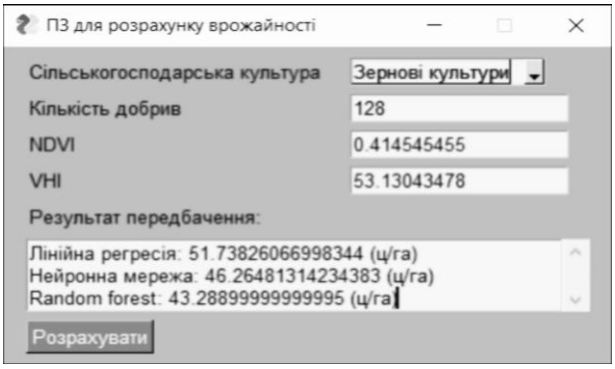

Рисунок 3.12 Результати прогнозування для зернових культур

Модель лінійної регресії передбачила врожайність на 2021 рік у 51,74 ц/га, що є завищеним прогнозом. Похибка прогнозування складає 9,2%.

Нейронна мережа передбачила врожайність на рівні 46,26 ц/га, що, порівняно з лінійною регресією, є трохи заниженим прогнозом. Похибка прогнозування складає 2,4%.

Модель на основі випадкового лісу передбачила врожайність у 43,29, що значно менше порівняно з реальним значенням. Похибка у прогнозі склала 8,7%.

Наступною сільськогосподарською культурою для передбачення є кормовий буряк. Реальний показник врожайності у 2021році склав 509 ц/га.

Результати прогнозування врожайності кормового буряку наведені на рис. 3.13.

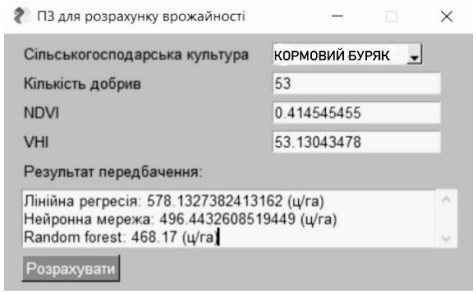

Рисунок 3.13 Результати прогнозування для кормового буряку

Модель лінійної регресії спрогнозувала врожайність у 578 ц/га, що значно більше за реальний показник. Похибка прогнозу склала 13,6%.

Нейронна мережа передбачила врожайність на рівні 496 ц/га, що доволі близько до реального значення. Похибка прогнозу склала 2,6%.

Модель на основі випадкового лісу передбачила врожайність у 468 ц/га. Похибка прогнозу склала 8,1%.

Останньою культурою для прогнозування є ріпак . Реальний показник врожайності у 2021році склав 23 ц/га.

Результати прогнозування врожайності ріпаку наведені на рис. 3.14.

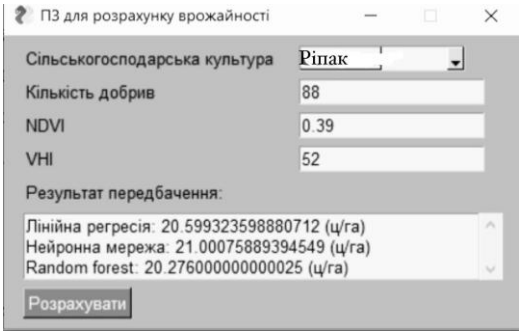

Рисунок 3.14 Результати прогнозування для ріпаку

Модель лінійної регресії передбачила врожайність у 24,9 ц/га. Похибка прогнозу склала 8,2%.

Нейронна мережа спрогнозувала врожайність на рівні 24,6 ц/га. Похибка прогнозу склала 7%.

Модель на основі випадкового лісу передбачила 19,5 ц/га, що значно нижче за реальний показник. Похибка у прогнозі склала 15,2%.

Отримані результати прогнозування врожайності занесемо до табл. 3.5.

Таблиця 3.5 Результати прогнозування врожайності

		Зернові культури	Кормовий буряк	Ріпак
	Реальне значення (ц/га)	47,6	508	26
Лінійна регресія	Передбачене значення (ц/га)	51,76	576	24,6
	Похибка (%)	9,5	13,4	8,4
Нейронна мережа	Передбачене значення (ц/га)	46,25	495	24,5
	Похибка (%)	2,1	2,4	7
Random forest	Передбачене значення (ц/га)	43,22	465	19,4
	Похибка (%)	8,7	8,1	15,2

Як можна побачити, найточніші результати дали моделі на основі нейронних мереж. Найбільша помилка склала 7%, що менше за мінімальні похибки в інших моделях. Середня похибка склала 4%.

Цікаво зазначити, що на етапі кросс-валідації (табл. 3.2-3.4) моделі на основі випадкового лісу дали найбільше значення MSE порівняно з моделями на інших методах, на зернових культурах та кормовому буряку, але одне з найменших на ріпаку. Проте на етапі експерименту вони дали протилежний результат: прогнозування зернових культур та кормового буряку дало меншу похибку, порівняно з моделями лінійної регресії, але модель для прогнозування врожайності ріпаку дала найбільшу похибку.

ВИСНОВКИ

У магістерській роботі подано теоретичне узагальнення та вирішення науково-прикладних проблем при використанні методів машинного навчання в прогнозуванні врожайності кількох важливих сільськогосподарських культур України.

На основі отриманих теоретичних і практичних результатів зроблено наступні висновки:

- Проведено аналіз літературних джерел щодо сучасного стану використання методів машинного навчання у прогнозуванні врожайності. На основі цього огляду було обґрунтовано актуальність обраної теми та визначено завдання.
- Описано кілька методів машинного навчання, які можна використовувати для вирішення завдань, а саме поліноміальну регресію, нейронні мережі та випадковий ліс. Для кожного методу були наведені математичні основи та описані найважливіші параметри та алгоритми, які використовуються.
- Розглянуто та описано різні типи вхідних даних, які можна використовувати для прогнозування врожайності сільськогосподарських культур. Особливу увагу приділили NDVI та VHI. Обґрунтовано доцільність відбору таких вхідних даних для вирішення проблеми.
- Для кожної культури, відібраної для дослідження, була створена вибірка з використанням описаних вхідних даних. Проведено попередній аналіз отриманих даних шляхом візуалізації та кореляційного аналізу. На основі цієї вибірки проведено оцінку точності передбачення деяких варіантів архітектури методів машинного навчання. Для отримання оцінки використано метод перехресної перевірки з MSE. На базі отриманої оцінки обґрунтовано вибір архітектури для створення кінцевого програмного забезпечення.

- Розроблено програмне забезпечення, яке використовує моделі машинного навчання для прогнозування врожайності сільськогосподарських культур в Україні. Для створення програми, аналізу даних та перехресної перевірки були використані мова високого рівня Python і набір бібліотек: Numpy, Seaborn, Pandas, Scikit-learn, Pickle, PySimpleGUI.
- Здійснено експериментальний прогноз урожайності сільськогосподарських культур на 2019-2020 рік. В результаті експерименту для кожного методу машинного навчання та культури була отримана похибка прогнозування. Середня похибка для лінійної регресії склала 10,3%, нейронної мережі – 4%, випадкового лісу – 10,6%.
- На основі аналізу отриманих експериментальних результатів показано, що для обраних рослин і вхідних даних найбільш оптимальними є моделі на основі нейронних мереж, які дали значно точніший результат, порівняно з іншими методами. Якщо вибрати інші вхідні дані, можна отримати інший результат, що підтверджується попередніми науковими публікаціями.
- Результати експерименту виявили розбіжності між оцінкою точності прогнозу перехресної перевірки та експерименту, що становить певний науковий інтерес.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Casey Quackenbush (2018). “A Painting Made by Artificial Intelligence Has Been Sold at Auction”. Time. : https://time.com/5435683/artificial-intelligence-painting-christies/

2. Google AI Blog. Show and Tell: image captioning open sourced in TensorFlow.: https://ai.googleblog.com/2016/09/show-and-tell-image-captioning-open.html

3. David Hest (2012). “New driverless tractor, grain cart systems coming this year”. Farm Industry News.: https://www.farmprogress.com/precision-guidance/new-driverless-tractor-grain-cart-systems-coming-year

4. Шумская Е. В. Прогнозирование урожайности зерновых культур на среднесрочный период.: Дис. ... канд. экономические науки: 08.00.12, 08.00.05, Москва. – 2007. – С.1-50.

5. Полевой А.Н., Русакова Т.И. и др. Прикладная динамическая модель формирования урожая сельскохозяйственных культур. // Сб. докладов: Гидрометеорологическое обеспечение агропромышленного комплекса страны. – Л.: Гидрометеоиздат. 1991. С. 5-30.

6. Просвиркина А.Г. Методы количественной оценки агрометеорологических условий формирования продуктивности и прогноза урожайности проса.: Дис. … канд. географические науки: 11.00.09, Москва. – 1984.

7. Игнатьев В.М. Моделирование урожайности сельскохозяйственных культур. Международный научно-исследовательский журнал.: https://research-journal.org/economical/modelirovanie-urozhajnosti-selskoxozyajstvennyx-kultur/

8. Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

9. Савин И.Ю., Барталев С.А., Лупян Е.А., Толпин В.А., Хвостиков СА. Прогнозирование урожайности сельскохозяйственных культур на основе

спутниковых данных: возможности и перспективы // Современные проблемы дистанционного зондирования Земли из космоса, 2010. Т.7. № 3. С. 275-285

10. Бекмуратов Т.Ф. Нечеткая модель прогнозирование урожайности / Мухамедиева Д.Т., Бобомурадов О.Ж. // Проблемы информатики. – 2010. – №3. – С. 11-23.

11. Темиров А.А. Алгоритм линейного клеточного автомата для прогнозирования урожайности зерновых / Новые технологии. – 2015. – №4.

12. Куссуль Н.М. Регресійні моделі прогнозування врожайності зернових в Україні за супутниковими даними різної природи / Колотій А.В., Яцків С.В., Олійник Т.В // Наукові праці Донецького національного технічного університету. Серія: Інформатика, кібернетика та обчислювальна техніка : збірник статей. Вип.1 (17) / ДВНЗ "ДонНТУ" ; редкол.: О.Є. Башков (голов. ред.) та ін. – Донецьк : ДонНТУ, 2013. Випуск 1 (17): http://ea.donntu.org:8080/jspui/handle/123456789/29679

13. Заводчиков Н.Д. Использование нейросетевых технологий в прогнозировании эффективности производства зерна / Спешилова Н.В., Таспаев С.С. // Известия Оренбургского государственного аграрного университета. – 2015. – №1 (51). – С. 216-219.

14. Хворова Л. А. Прогнозирование урожайности зерновых культур: методы и расчеты / Гавриловская Н. В. // Известия Алтайского государственного университета. – 2008. – №1. – С. 65-68.

15. Гагарин А.Г. Прогнозирование урожайности на основе анализа кросс-региональных данных / Рогачев А.Ф. // Известия Нижневолжского агроуниверситетского комплекса: наука и высшее профессиональное образование. – 2018. – №2 (50). – С. 339-345.

16. Розенберг Г.С. Екологічне прогнозування (функціональні предиктори часових рядів). / Шитіков В.К., Брусіловский П.М. – Тольятті, 1994. 182 с.

17. Samuel, Arthur (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development (3): 210–229.

18. Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer.

19. Alpaydin, Ethem (2010). Introduction to Machine Learning. London: The MIT Press.

20. Sarle, Warren (1994). "Neural Networks and statistical models".

21. Langley, Pat (2011). "The changing science of machine learning".

22. Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall.

23. Львівський національний університет імені Івана Франка. Машинне навчання простими словами. Частина 1. – Режим доступу до електронного ресурсу: http://www.mmf.lnu.edu.ua/ar/1739

24. David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press.

25. Towards data science. “Overfitting vs. Underfitting: A Complete Example”: https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765

26. Hilary L. Seal (1967). "The historical development of the Gauss linear model". Biometrika.

27. Scikit-learn. “Underfitting vs. Overfitting”: https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

28. The Asimov Institute. The neural network zoo.: https://www.asimovinstitute.org/neural-network-zoo/

29. Kleene, S.C. (1956). "Representation of Events in Nerve Nets and Finite Automata". Princeton University Press.

30. Cybenko, G.V. (1989). "Approximation by Superpositions of a Sigmoidal function". Mathematics of Control, Signals and Systems.

31. Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. С. 278–282.

32. Паклин Н., Орешков В. Бизнес-аналитика: от данных к знаниям. 2-е издание. Питер 2013. С. 428-433.

33. Интуит. «Методы классификации и прогнозирования. Деревья решений»: https://www.intuit.ru/studies/courses/6/6/lecture/174

34. Rokach L. Ensemble-based classifiers // Artificial Intelligence Review. – 2010. – Т. 33, вип. 1-2. С. 1-39.

35. Condorcet N. C. Essai sur l'application de l'analyse à la Probabilité des Décisions rendues a la Pluralité des voix. Paris: L'Imprimerie Royale, 1785.

36. Breiman, Leo (September 1994). "Bagging Predictors". Department of Statistics, University of California Berkeley.

37. Machine Learning Demystified. "Bagging and Boosting": https://prachimjoshi.wordpress.com/2015/07/23/bagging-and-boosting/

38. Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests". IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8), 832–844.

39. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer, 592

40. Посудін Ю. І. Методи вимірювання параметрів навколишнього середовища: Підручник. – К.: Світ (видавництво), 2003. – 288 с.

41. Lillesand, T.M., Kiefer, R.W. and Chipman, J.W. (2004) Remote Sensing and Image Interpretation. 5th Edition, John Wiley & Sons Ltd., Hoboken.

42. NOAA STAR Center for satellite application and research: https://www.star.nesdis.noaa.gov/

43. Earth Observing System (EOS): NDVI FAQ: All you need to know about NDVI (2009). Режим доступу до електронного ресурсу: https://eos.com/blog/ndvi-faq-all-you-need-to-know-about-ndvi/

44. ResearchGate. Режим доступу до електронного ресурсу: https://www.researchgate.net/figure/NDVI-seasonal-profile-NDVIm-is-the-annual-mean-of-NDVI-and-a-surrogate-of-annual-primary_fig1_314294912

45. Kogan F.N., Global Drought Watch from Space. Bull. Am. Meteorol. Soc. 1997, 78, 621–636.

46. Kogan F.N., Operational space technology for global vegetation assessment. Bull. Am. Meteorol. Soc. 2001, 82, 1949–1964.

47. Kogan F.N., “Application of vegetation index and brightness temperature for drought detection,” Adv. Space Res. 15, pp. 91–100, 1995.

48. Bento, V.A.; Gouveia, C.M.; DaCamara, C.C.; Trigo, I.F. A climatological assessment of drought impact on vegetation health index. Agric. For. Meteorol. 2018, 259, 286–295.

49. A. Karnieli et al., “Use of NDVI and Land Surface Temperature for Drought Assessment : Merits and Limitations,” JOURNAL OF CLIMATE, vol. 23, pp. 618–633, 2009

50. Kogan F., Salazar L., Roytman L. Forecasting crop production using satellite-based vegetation health indices in Kansas, USA // International Journal of Remote Sensing. – 2012. – 33, N 9. – P. 2798-2814.

51. Сайт державної служби статистики України: http://www.ukrstat.gov.ua/

Додаток А

Лістинг програмного застосунку

```
import PySimpleGUI as sg
import pickle
import sklearn

model_lz = pickle.load(open('models\model_lz.sav', 'rb'))
model_lb = pickle.load(open('models\model_lb.sav', 'rb'))
model_ls = pickle.load(open('models\model_ls.sav', 'rb'))
model_nz = pickle.load(open('models\model_nz.sav', 'rb'))
model_nb = pickle.load(open('models\model_nb.sav', 'rb'))
model_ns = pickle.load(open('models\model_ns.sav', 'rb'))
model_fz = pickle.load(open('models\model_fz.sav', 'rb'))
model_fb = pickle.load(open('models\model_fb.sav', 'rb'))
model_fs = pickle.load(open('models\model_fs.sav', 'rb'))

sg.theme('Light Blue 2')

layout = [ [sg.Text('Сільськогосподарська культура', size=(25, 1)),sg.Drop(key = 'type', values=('Зернові культури', 'Цукровий буряк', 'Соняшник'))],
           [sg.Text('Кількість добрив', size=(25, 1)), sg.InputText(size=(20, 1), key = 'fert')],
           [sg.Text('NDVI', size=(25, 1)), sg.InputText(size=(20, 1), key = 'NDVI')],
           [sg.Text('VHI', size=(25, 1)), sg.InputText(size=(20, 1), key = 'VHI')],
           [sg.Text('Результат передбачення:')],
           [sg.Output(size=(50,3), key='res')],
```

```
            [sg.Button('Розрахувати')] ]

    window = sg.Window('ПЗ для розрахунку врожайності', layout)

    while True:
        event, values = window.read()

        if event == 'Розрахувати':
            if values['fert'] == '' or values['NDVI'] == '' or values['VHI'] == '' or values['type'] == '':
                sg.popup('Усі дані повинні бути введені!')
            elif values['type'] == 'Зернові культури':
                try:
                    data = [[float(values['fert']), float(values['NDVI']), float(values['VHI'])]]
                except:
                    sg.popup('Дані введено некоректно!')
                    continue
                text = 'Лінійна регресія: ' + str(model_lz.predict(data)[0]) + ' (ц/га)\nНейронна мережа: ' + str(model_nz.predict(data)[0]) + ' (ц/га)\nRandom forest: ' + str(model_fz.predict(data)[0]) + ' (ц/га)'
                window['res'].update(text)

            elif values['type'] == 'Кормовий буряк':
                try:
                    data = [[float(values['fert']), float(values['NDVI']), float(values['VHI'])]]
                except:
                    sg.popup('Дані введено некоректно!')
                    continue
```

```
            text = 'Лінійна регресія: ' + str(model_lb.predict(data)[0]) + ' (ц/га)\nНейронна мережа: ' + str(model_nb.predict(data)[0]) + ' (ц/га)\nRandom forest: ' + str(model_fb.predict(data)[0]) + ' (ц/га)'
            window['res'].update(text)

        elif values['type'] == 'Ріпак':
            try:
                data = [[float(values['fert']), float(values['NDVI']), float(values['VHI'])]]
            except:
                sg.popup('Дані введено некоректно!')
                continue
            text = 'Лінійна регресія: ' + str(model_ls.predict(data)[0]) + ' (ц/га)\nНейронна мережа: ' + str(model_ns.predict(data)[0]) + ' (ц/га)\nRandom forest: ' + str(model_fs.predict(data)[0]) + ' (ц/га)'
            window['res'].update(text)

    window.close()
```