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

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня «магістр»

НА ТЕМУ:

Методи машинного навчання в прогнозуванні врожайності сільськогосподарських культур

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

Виконав:

студент 2 курсу магістратури
групи ТШІ-21
Бородай Денис Юрійович

Науковий керівник:

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

Кваліфікаційна робота магістра допущена до захисту
рішенням кафедри *інтелектуальних технологій*

Протокол № ____ від «____» травня 2020 р.

В.о. зав. кафедри _____ доц. Красовська Г.В.
(підпис)

Київ 2020

РЕФЕРАТ

Кваліфікаційна робота складається зі вступу, 3 розділів, висновків, списку використаної літератури із 52 джерел та 1 додатку. Загальний обсяг роботи 79 сторінок. Робота містить 6 таблиць та 34 рисунки.

Актуальність теми. Проблема математичного моделювання та прогнозування врожайності сільськогосподарських культур є важливою для планування аграрного виробництва. Водночас вплив ряду природно-кліматичних, біологічних та організаційно-технологічних груп чинників надає різноспрямований вплив на результати прогнозування, приводячи до неприпустимо високої, більш ніж 15%, похибки. Застосування класичних підходів математичного моделювання, таких як побудова систем рівнянь економетрики, різного виду адаптивних моделей, а також сучасних методів нелінійної динаміки, не завжди призводить 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.

Метою кваліфікаційної роботи є дослідження ефективності методів машинного навчання в 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.

Результати роботи. У ході роботи було описано декілька методів машинного навчання, на основі яких розроблено програмний додаток, який використовує моделі на основі цих методів для прогнозування врожайності сільськогосподарських культур на території України. Використовуючи

розроблений додаток, було проведено експериментальне прогнозування врожайності сільськогосподарських культур на 2018 рік. У результаті експерименту було отримано похибку прогнозу для кожного методу машинного навчання та кожної сільськогосподарської культури. На основі аналізу отриманих експериментальних результатів показано, що для обраних сільськогосподарських культур та вхідних даних найоптимальнішими являються моделі на основі нейронних мереж, які дали набагато точніший результат порівняно з іншими методами.

Апробація результатів кваліфікаційної роботи. Основні положення кваліфікаційної роботи доповідалися на конференціях:

1) V Міжнародна науково-практична конференція «Обчислювальний інтелект» (м. Ужгород, 15-20 квітня 2019 р.);
2) VI Міжнародна науково-практична конференція «Інформаційні технології та взаємодії» (м. Київ, 20 грудня 2019 р.).

За матеріалами кваліфікаційної роботи опубліковано 2 публікації у збірниках матеріалів цих науково-практичних конференцій.

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, list of references (52 sources) and 1 application. The total volume of work is 79 pages. The work contains 6 tables and 49 figures.

Actuality of theme. The problem of mathematical modeling and forecasting of crop yields is important for agricultural production planning. At the same time, the influence of a number of natural-climatic, biological and organizational-technological factors has a multidirectional influence on the forecasting results, leading to an unacceptably high error, more than 15%. In addition, classical mathematical modeling, such as systems of equations of econometrics, various types of adaptive models or modern methods of nonlinear dynamics doesn't always lead to good results. Therefore, using and researching new classes of mathematical models in agriculture, such as models based on machine learning methods, is an actual and perspective direction.

The purpose of the qualification work is to research the effectiveness of machine learning methods in forecasting crop yields.

The object of research is prediction crop yields using machine learning methods.

The subject of research is systems for forecasting crop yields based on different machine learning methods.

Results of the work. In qualification work were described several machine learning methods. A software application was developed using models based on these methods for forecasting crop yields on the territory of Ukraine. Using the developed application, was conducted an experimental forecast of crop yields for 2018. As a result of the experiment, a forecast error was obtained for each machine learning method and each crop. Based on the analysis of experimental results, it is shown that for the selected

crops and input data, the models based on neural networks are the most optimal, which gave a much more accurate result compared to other methods.

Publications. The main results of the qualification work were presented as thesis at conferences:

1) V International Scientific and Practical Conference "Computational Intelligence" (Uzhgorod, April 15-20, 2019);
2) VI International Scientific and Practical Conference "Information Technology and Interaction" (Kyiv, December 20, 2019).

Based on the materials of qualification work, 2 scientific papers were published in the scientific and practical conferences.

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

ЗМІСТ

ВСТУП

Основна задача науки та реального життя – отримання достовірних прогнозів про майбутню поведінку складних систем, об’єктів або явищ на підставі їх минулої поведінки. Прогнозування є важливим компонентом сучасних інформаційних технологій прийняття рішень при проектуванні складних систем (паливно-енергетичних, агротехнічних, інформаційно-комунікаційних тощо) й управління ними в умовах невизначеності. Ефективність того чи іншого рішення оцінюється на основі подій та результатів, які виникають вже після його прийняття та реалізації. Тому прогнозування та оцінка наслідків реалізації альтернатив рішень, що приймаються на етапі їх формування й аналізу, дозволяють здійснити більш правильний вибір рішення і значно знизити ризики настання несприятливих наслідків.

Багато задач, що виникають на практиці, не можуть бути вирішені заздалегідь відомими методами або алгоритмами. Це відбувається з тієї причини, що нам заздалегідь невідомі механізми походження вихідних даних, або ж відома нам інформація є недостатньою для побудови прогнозної моделі. Таким чином, ми отримуємо дані з «чорного ящика». У цих умовах нічого не залишається, як тільки вивчати доступну нам послідовність вихідних даних і намагатися побудувати модель передбачення й удосконалювати її в процесі роботи. Підхід, при якому минулі дані або приклади використовуються для початкового формування та вдосконалення моделі передбачення, називається методом машинного навчання.

Перш за все, методи, які базуються на машинному навчанні, стають популярними завдяки високій ефективності та можливості обробляти величезні обсяги отриманої інформації. Машинне навчання – надзвичайно широка область досліджень, що динамічно розвивається, і яка використовує величезне число теоретичних і практичних методів.

На сьогодні, майже в усіх сферах використовується машинне навчання, починаючи від звичайного запиту в пошуковій системі, до безпілотного автономного транспорту. Така б, здавалося, традиційна сфера, як сільське господарство не є виключенням. З кожним роком широкого застосування у аграрній сфері набувають саме методи прогнозування та покращення врожайності, які базуються на машинному навчанні. Ця робота й присвячена саме таким методам. Нині вже такі технології використовують як великі компанії та холдинги, так і невеликі стартапи.

Проблема математичного моделювання та точного прогнозування врожайності сільськогосподарських культур є важливою для планування аграрного виробництва. Водночас вплив ряду природно-кліматичних, біологічних та організаційно-технологічних груп чинників надає різноспрямований вплив на результати прогнозування, приводячи до неприпустимо високої, більш ніж 15%, похибки. Застосування класичних підходів математичного моделювання, таких як побудова систем рівнянь економетрики, різного виду адаптивних моделей, а також сучасних методів нелінійної динаміки, не завжди призводить 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 магістерській роботі ми ознайомимося з деякими сучасними математичними проблемами даної галузі та їх вирішенням, основною з яких є проблема побудови моделей і оцінка якості їх передбачень.

Метою магістерської роботи є дослідження ефективності методів машинного навчання в 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 досліджень є:

1. провести аналіз літературних та інших джерел, зокрема Інтернет, щодо прогнозування врожайності сільськогосподарських культур;
2. описати концепцію точного землеробства та її вплив на збільшення врожайності сільськогосподарських культур;
3. розглянути різні підходи та методи машинного навчання для побудови систем прогнозування;
4. обрати й описати вхідні дані та провести їх аналіз;
5. обрати необхідну архітектуру для кожної моделі машинного навчання й обґрунтувати її вибір;
6. розробити систему прогнозування врожайності сільськогосподарських культур на основі різних методів машинного навчання;
7. здійснити опис та аналіз отриманих результатів;
8. встановити придатність та 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

Впродовж останніх десятиліть відбувся ривок в обчислювальних та інформаційних технологіях. Завдяки йому обробляються величезні об’єми даних у різних областях, таких як медицина, біологія, фінанси та маркетинг тощо. Задача розуміння цих даних привела до появи нових інструментів в області статистики та породила нові сфери, такі як інтелектуальний аналіз даних (дейта майнінг), машинне навчання та біоінформатика.

На сьогодні машинне навчання продовжує набирати популярність і поступово входить у всі аспекти повсякденного життя. Провідні виробники зробили машинне навчання невід’ємною частиною смартфонів, мобільних додатків, побутової техніки, щоб забезпечити більш ефективне користування ними. Наразі не існує майже проблем, які не намагалися б вирішити за допомогою машинного навчання, починаючи від знаходження помилок у тексті та закінчуючи створенням витворів мистецтва [1].

Методи машинного навчання набули своєї популярності завдяки високій ефективності та можливості обробляти велику кількість інформації, яку на цей час можна отримати. Так, наприклад, сучасні системи розпізнавання зображень досягають точності в 93,9% [2].

Така б, здавалося, традиційна сфера, як сільське господарство, не є виключенням й активно застосовує машинне навчання в останні роки. В той час, коли автовиробники тільки тестують функції автопілоту, виробники тракторів вже як декілька років продають автономні трактори [3].

В сучасних умовах ринкових відносин та самостійності сільськогосподарських угідь, питання передбачення об’ємів врожайності набуває все більшої актуальності.

Серед багатьох показників, що описують діяльність сільського господарства, особливої уваги заслуговують техніко-економічні показники та показники врожайності сільськогосподарських культур, які є комплексними показниками. Дані показники, з одного боку, являються є цінною інформацією для побудови прогнозів, планів та допомагають у прийнятті управлінських рішень, а з іншої сторони можуть використовуватися як оцінка ефективності сільського господарства [4].

Врожайність сільськогосподарських культур є складним показником, з точки зору моделювання, оскільки формування врожаю залежить не тільки від виробничих факторів, а й від біологічних та погодних показників. Тому показники, які впливають на врожайність сільськогосподарських культур можна розділити на дві групи: технологічний рівень землеробства та погодні фактори [5].

Отримання точного прогнозу врожайності дозволяє коректно розв’язувати питання формування резервних продовольчих фондів та будувати ефективну й адекватну політику зовнішньої торгівлі [4].

Таким чином, це робить високопріоритетною задачу методології побудови систем комплексного агрометеорологічного обслуговування сільського господарства, яка має єдину методичну основу для всіх складових системи: як для набору культур, так і для методів пронозу й оцінки умов формування врожаю. Такою методичною основою є математичні моделі [6].

Математичні моделі – це опис деякого об’єкта або явища на основі математичного апарату (рис. 1.1). Вони використовуються для дослідження систем та впливів різних параметрів на їх поведінку. Особливо корисні математичні моделі у тому випадку, коли з реальним об’єктом або явищем

складно експериментувати, наприклад, якщо явище рідкісне, чи занадто дорого обходиться експеримент.

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

Рисунок 1.1 Приклад математичного моделювання врожайності сільськогосподарських культур [7]

В області математичного моделювання врожайності досягнуто значних успіхів: розроблено моделі основних процесів життєдіяльності, які впливають на продуктивність рослинності, та моделі формування врожайності культур, призначені для прогнозування, планування та управління в сільському господарстві [6].

В останні десятиліття все більшою популярністю для моделювання врожайності сільськогосподарських культур користуються методи машинного навчання, які є свого роду еволюцією математичного моделювання.

У першу чергу, це зумовлено ростом обчислювальних можливостей та їх доступністю. Крім цього, значно зросла кількість доступних наборів даних. Так за прогнозами до 2025 року порівняно з 2019 їх кількість збільшиться більш ніж у 4 рази (рис. 1.2) [8].

Варто також зазначити, значний зріст кількості доступних супутникових даних, який, за останні роки, зумовив пришвидшений розвиток методів та інформаційних технологій прогнозування врожайності на основі супутникової інформації [9].

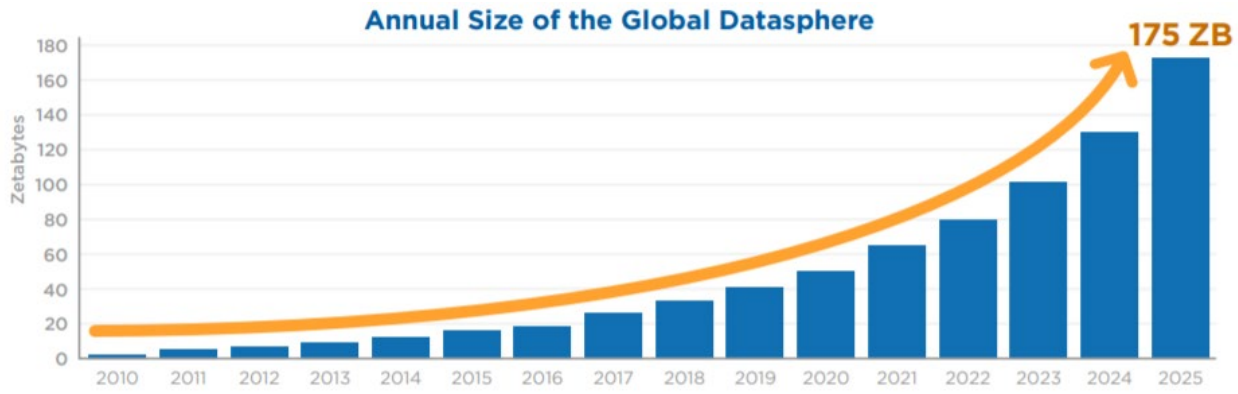

Рисунок 1.2 Кількість даних у світі з прогнозом до 2025 року [8]

Існує значна кількість наукових робіт, пов’язаних з використанням методів машинного навчання для прогнозування врожайності сільськогосподарських культур як серед вітчизняних, так і серед іноземних науковців.

Так, наприклад, у роботі Т.Ф. Бекмуратова, Д. Т. Мухамедієва, О.Ж. Бобомурадова «Нечітка модель прогнозування врожайності» [10] використовувалися нечітко-множинні моделі типу Сугено для прогнозування врожайності бавовни різних селекційних сортів в умовах нечітко заданих типів ґрунту, режимів поливу та внесення добрив, а також погодних умов. У роботі застосовується аналітичне моделювання прогнозу врожайності та аналізуються аналітичні залежності прогнозної врожайності залежно від різних факторів В якості вхідних параметрів використовувалися такі нечіткі дані: погодні умови при посіві, водозабезпеченість, погодні умови при вегетації, погодні умови при зборі врожаю, тип ґрунту, тип селекційного сорту, режим внесення добрив. Результати експерименту показали високу ефективність прогнозування на основі моделей Сугено, зокрема похибка прогнозу склала 0÷2,77%.

У роботі А.А. Темірова «Алгоритм лінійного клітинного автомату для прогнозування врожайності зернових» [11] використовувався алгоритм лінійного клітинного автомату для побудови моделі для прогнозування врожайності зернових на основі часових рядів. У роботі застосовувалися моделі прогнозування на основі клітинних автоматів з різними конфігураціями. Для моделювання структури клітинного автомату та його навчання використовувався

генетичний алгоритм. У якості вхідних даних використовуються часові ряди щорічної врожайності зернових з 1896 по 2014 рр. У ході роботи було підтверджено адекватність клітинно-автоматної прогнозної моделі для прогнозування числового ряду врожайності зернових. Похибка прогнозу склала 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] використовувалися регресійні моделі прогнозування врожайності зернових культур на основі супутникових даних. У роботі проводиться порівняльний аналіз використання різних супутникових даних для прогнозування врожайності озимої пшениці в Україні для різних областей. Аналіз результатів показав, що використання індексу NDVI дає більшу похибку прогнозу, порівняно з FAPAR і VHI, але обмежений обсяг даних не дозволяє стверджувати, що NDVI завжди буде забезпечувати менш точний прогноз. Тому авторам видається доцільним не обмежуватися однією прогнозною моделлю, а використовувати інші підходи до побудови моделей на основі предикторів різної природи.

У роботі Н.Д. Заводчікова, Н.В. Впешилова, С.С. Таспаєва «Використання нейромережевих технологій у прогнозуванні ефективності виробництва зерна» [13] застосовується імітаційне моделювання на основі нейронної мережі для прогнозування врожайності зернових культур та проводиться розрахунок врожайності при зміні деяких параметрів на основі отриманої моделі. У якості вхідних даних використовувалися 9 показників, які, згідно з кореляційним аналізом, мали найбільший вплив на врожайність. Результати роботи підтвердили доцільність використання нейронних мереж для вирішення задачі оптимізації виробництва зерна.

У роботі Л.А. Хворової, Н.В. Гавриловской «Прогнозування врожайності зернових культур: методи та розрахунки» [14] визначався рік-аналог за допомогою методів класифікації та розпізнавання образів з метою здійснення

прогнозу врожайності ярої пшениці. Для створення моделі проводиться кластеризація років за певними параметрами, а для створення прогнозу виконується класифікація року на основі отриманих кластерів у моделі. На основі дисперсійного аналізу з усіх показників були обрані у якості параметрів наступні: сума ефективних температур, сума опадів, кількість днів з опадами, дефіцит вологості насичення. Результатом роботи стала модель, яка дозволяє взяти погодний сценарій всього вегетаційного періоду року-аналогу та здійснити уточнювальний прогноз врожайності на поточний рік. Використовуючи дані по рокам-аналогам можна здійснювати попередній прогноз вже після двох-трьох тижнів вегетаційного періоду. Автори зазначають, що проведену класифікацію та прогноз врожайності ярої пшениці слід розглядати як початковий етап роботи по оцінці врожайності зернових культур. Однак це показує, що дана метрика має цілком певний сенс і достатньо добре дозволяє здійснювати прогноз.

У роботі А.Г. Гагаріна, А.Ф. Рогачєва «Прогнозування врожайності на основі аналізу кросс-регіональних даних» [15] використовувалися нейронні мережі для побудови моделей прогнозування врожайності озимої пшениці. Для навчання застосовувалася вибірка, яка складається з 54 часових рядів зі значеннями врожайності озимої пшениці за 21 рік. На основі проведених прогнозних експериментів та аналізу статистичних характеристик часових рядів врожайності досліджуваних сільськогосподарських культур, було обґрунтовано вибір програмних засобів, структуру штучної нейронної мережі, виконано її навчання та доведено можливість отримання короткострокових прогнозів з похибкою в межах 15-20%. Автори зазначають, що використання розроблених генеративних змагальних нейронних систем забезпечує розв’язання задачі прогнозування врожайності на прикладі зернових культур на основі крос-регіонального аналізу економічних та кліматологічних даних.

При аналізі цих та подібних робіт можна зазначити, що в них використовуються різні методи машинного навчання, типи вхідних даних, сільськогосподарські культури, терміни прогнозування тощо. На базі цього

можна зробити висновок, що на цей час існує значний простір для досліджень у даному напрямку й нестача систематичних досліджень щодо ефективності різних параметрів.

1.2 Аналіз існуючих систем прогнозування врожайності

Недивлячись на велику кількість наукових робіт на дану тему з доволі точними результатами, нині не існує окремих комерційних систем для прогнозування врожайності. Проте в останні роки широкого розповсюдження набрала концепція точного землеробства.

Точне землеробство – це концепція з використання комплексних систем на основі різних технологічних рішень, які можуть збільшити врожайність й допомогти краще управляти аграрними ресурсами.

Для точного землеробства використовується цілий ряд технологій та показників:

- геоінформаційні системи (ГІС);
- системи глобального позиціонування (GPS);
- дистанційне зондування Землі (ДЗЗ);
- технології змінного нормування (Variable Rate Technology);
- дані зі спеціальних датчиків;
- тощо.

В основі концепції точного землеробства лежить те, що умови для розвитку рослин в різних місцях одного й того ж самого поля подібні, проте неоднакові. Таким чином, існують неоднорідності у межах одного поля (рис. 1.3).

Рисунок 1.3 Приклад неоднорідності на одному полі

Точне землеробство базується на використанні максимально деталізованих по характеристикам карт конкретних полів. Наявні кадастрові карти надають мало корисної інформації, а саме лише кордони полів. Крім цієї інформації, необхідно мати дані по рівню вологості ґрунту, хімічному складу, куту нахилу поверхні, кількості сонячного випромінювання тощо. Чим більше факторів містить така карта, тим точніше будуть працювати комп'ютерні та супутникові технології, й тим швидше та ефективніше можна буде вносити корективи у виробничий процес.

Системи точного землеробства часто являють собою системи підтримки рішень, тому на основі отриманих карт надаються точні рекомендації. Для кожної ділянки розраховується необхідна кількість, води, насіння, добрив.

На основі рекомендацій формуються інструкції, які завантажуються на бортовий комп'ютер сільськогосподарської техніки. Під час виконання інструкцій, людина лише контролює правильність їх виконання машиною. Техніка, на основі супутникової навігації, рухається полем та вносить насіння й

добрива, регулюючи їх кількість на конкретній ділянці поля згідно з отриманими інструкціями. За рахунок використання GPS шлях прокладається таким чином, аби не було нашарування чи прогалин між обробленими смугами.

Отже, використовуючи точне землеробство, можна отримати переваги одразу по декільком напрямкам:

- значно зменшуються витрати насіння та матеріалів, і, як результат, зменшується собівартість продукції;
- збільшується врожайність й прибуток;
- отримується продукція вищої якості;
- покращується якість ґрунту;
- зменшується негативний вплив на навколишнє середовище, шляхом зменшення кількості внесених добрив.

Проаналізувавши системи точного землеробства, можна зазначити, що є можливість отримання таких параметрів поля:

- просторові дані;
- карти врожайності;
- фотознімки та карти NDVI;
- дані польового обладнання;
- стан ґрунту;
- дані про внесення добрив;
- метеорологічні дані;
- тощо.

На основі такого набору даних можна побудувати доволі точні моделі для прогнозування врожайності сільськогосподарських культур для кожного окремого поля. Крім того, більшість систем точного землеробства мають архіви показників за останній десяток років, що ще більше спрощує задачу. Проте, невідомо з яких причин, до цього часу у більшості таких систем відсутня можливість прогнозування врожайності. Тому, беручи до уваги розглянуту

актуальність даного питання, було б доцільно додати можливість прогнозування врожайності у системах точного землеробства.

1.3 Постановка задачі

Майже в усіх розглянутих вище роботах використовуються різні методи для побудови систем прогнозування. У деяких робиться розгляд впливу різних вхідних даних на результати роботи системи, проте не розглядається порівняння моделей на основі різних методів машинного навчання.

Г.С. Розенберг зі співавторами [16] вказує на те, що специфіка прогнозування на сучасному етапі полягає, перш за все, у баченні одного і того ж феномену за допомогою декількох різних і більш-менш рівноцінних моделей (прояв принципу множинності моделей). На їхню думку, основний недолік наявних систем прогнозування полягає в тому, що прогноз конкретного часового ряду будується в рамках тільки одного алгоритму. Інакше кажучи, передбачається, що істинний механізм генерації цього ряду є єдиним, і що він добре апроксимується одним з алгоритмів.

Тому залишається невирішеним питання: система на основі якого методу буде давати більш точний результат.

Дане питання є надзвичайно важливим, оскільки різниця у точності прогнозу навіть на долі відсотка можуть мати значний вплив на кількість врожаю й, як результат, на дохід аграріїв та країни. Тому у цій роботі розглядається побудова системи прогнозування врожайності сільськогосподарських культур за допомогою 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 підгалузь штучного інтелекту про алгоритми та статистичні моделі, які використовують комп'ютерні системи для ефективного виконання специфічних задач без використання чітких інструкцій, опираючись замість них на шаблони та логічні висновки. Алгоритми машинного навчання будують математичні моделі на основі вибірок даних для того, щоб робити передбачення або висновки без явного програмування для виконання заданої цілі [17, 18].

Основною ціллю системи, що навчається, є одержання висновків (узагальнень) на основі свого досвіду [18]. Узагальнення, в даному випадку, – це здатність цієї системи точно виконувати нові, досі невирішені завдання, після отримання досвіду на основі навчальної вибірки. На основі навчальної вибірки система повинна побудувати загальну модель, завдяки якій вона буде здійснювати передбачення у нових випадках [19].

Термін «машинне навчання» було введено у 1959 році Самюелем Артуром [17]. Ще у перших роботах зі створення штучного інтелекту, деякі науковці були зацікавлені у машинах, які навчаються на основі даних. Вони намагалися досягнути цієї мети завдяки різним символьним методам. В основному це були перцептрони та інші моделі, в основі яких були узагальнені лінійні статистичні моделі, які згодом були названі «нейромережами» [20].

На початку 80-х років з'явилися експертні системи, які почали домінувати у сфері штучного інтелекту [21]. Завдяки цьому, а також з недостатністю обчислювальних можливостей та проблемами з накопиченням даних, відбувся

занепад машинного навчання (яке, на той час, ще не було відокремлене від штучного інтелекту) [22].

У 90-х роках відбулося відродження машинного навчання, яке було виокремлене в окрему підгалузь штучного інтелекту. Основною метою нової підгалузі стало вирішення задач практичного характеру. В її основі почали лежати методи статистики та теорії ймовірностей, на відміну від символьних методів, які використовувалися в штучному інтелекті [21]. Крім того, значною мірою зросли обчислювальні можливості та завдяки розвитку Інтернету відбулося спрощення у доступі до даних, що, своєю чергою, сприяло розвитку машинного навчання.

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

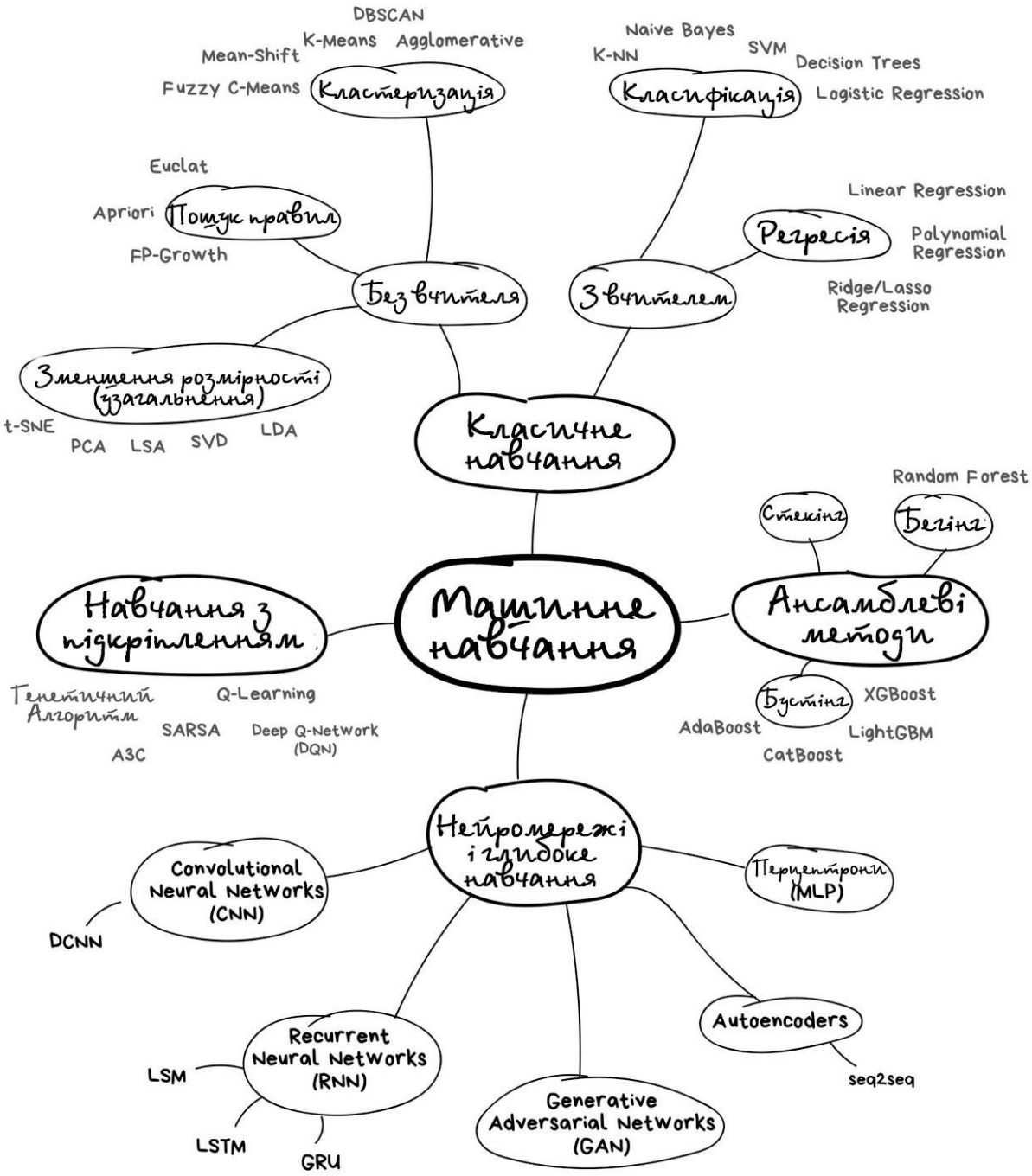

Рисунок 2.1 Загальна класифікація алгоритмів машинного навчання [23]

Існує декілька способів (типів) машинного навчання: навчання з вчителем, навчання без вчителя та навчання з підкріпленням. Розглянемо детальніше кожен з них та задачі, які вони розв’язують.

Навчання з вчителем – спосіб машинного навчання, при якому усі дані розмічені. Під час навчання результат моделі порівнюється з правильною відповіддю й на основі цього відбувається навчання. Даний спосіб машинного навчання розв’язує такі задачі:

- регресійні задачі – на основі вхідних даних отримати скаляр або вектор;
- задачі класифікації – на основі вхідних даних визначити до якого класу віднести об’єкт.

Навчання без вчителя – спосіб машинного навчання, при якому дані не розмічені. Під час навчання моделі повинні самі знаходити закономірності й робити висновки на основі даних. Даний спосіб розв’язує такі задачі:

- кластеризація – пошук подібних об’єктів, об’єднання їх у класи та визначення самих класів;
- зниження розмірності даних – зниження розмірності даних шляхом збору певних ознак у абстракції вищого рівня;
- пошук правил (асоціацій) – пошук закономірностей або правил на основі даних.

Навчання з підкріпленням – спосіб машинного навчання, при якому система (агент) взаємодіє з певним середовищем. Мета такого навчання мінімізувати помилку, а не запам’ятати або розрахувати усі можливі варіанти. Прикладом використання може бути автопілот або штучний інтелект в іграх.

Таким чином, для поставленої задачі підходить навчання з вчителем, оскільки необхідно розв’язати саме регресійну задачу: на основі вхідних даних передбачити числовий результат (врожайність). Розглянемо підходящі методи машинного навчання, які підходять для такої задачі.

2.1.2 Лінійна та поліноміальна регресія

Лінійна регресія – це лінійний підхід у моделювання зв'язків між скалярними залежними та векторами незалежних змінних (рис. 2.2). У випадку з однією залежною змінною вона має назву проста лінійна регресія. Якщо залежних змінних декілька – множинна лінійна регресія [24]. Відповідно до назви метод вирішує задачі регресії.

Рисунок 2.2 Приклад лінійної регресії [25]

У лінійній регресії зв'язки моделюються за допомогою лінійної предикативної функції, вектор параметрів якої невідомі й розраховуються з даних. Такі моделі називаються лінійними [26].

У загальному випадку модель лінійної регресії має наступний вигляд:

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

де

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

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

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

θ_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:

$$y = X\theta + \varepsilon,$$

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

$X = \begin{pmatrix} x_1^T \, x_2^T \\ \vdots \, x_n^T \end{pmatrix} = \begin{pmatrix} 1 \, x_{11} \, 1 \, x_{21} & \dots \, x_{1p} \, \dots \, x_{2p} \\ \vdots & \vdots \\ 1 \, x_{n1} & \dots \, x_{np} \end{pmatrix}$ – матриця векторів незалежної змінної (усіх *i*-их випадків);

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

$\varepsilon = \begin{pmatrix} \varepsilon_1 \, \varepsilon_2 \\ \vdots \, \varepsilon_n \end{pmatrix}$ – 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).

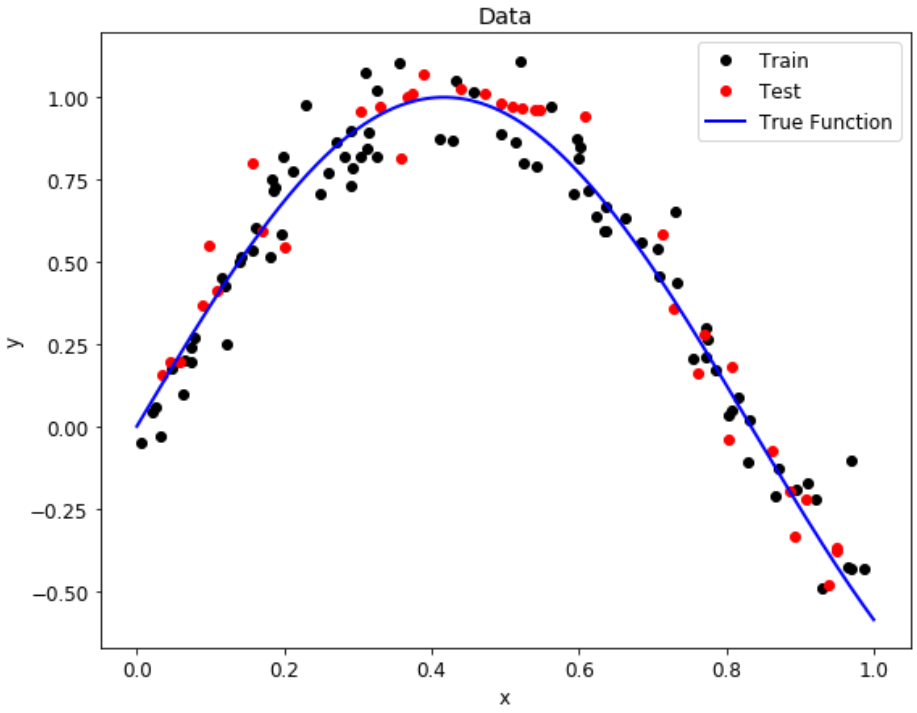

Рисунок 2.3 Приклад поліноміальної регресії [25]

При використанні поліноміальної регресії необхідно використати поліном такої степені, який найкраще зможе апроксимувати шукану залежність. Для знаходження оптимальної степені використовується кросс-валідація.

Якщо використати поліном занадто низької степені, то модель буде занадто простою (underfitting) і їй не вдасться правильно віднайти залежність (рис. 2.4). При занадто простій моделі помилка буде високою як на тренувальній, так і на тестовій вибірці.

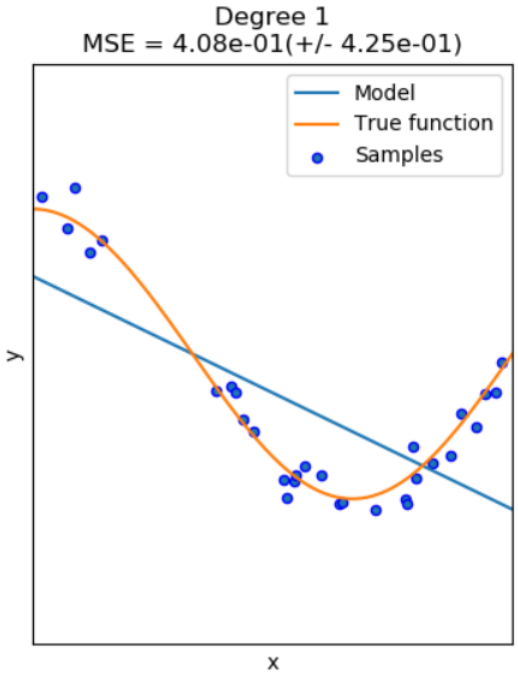

Рисунок 2.4 Занадто проста поліноміальна модель (underfitting) [27]

Якщо ж використати зависоку степінь в поліномі, то модель буде перенавченою (overfitting) й буде намагатися максимально «запам'ятовувати» приклади з навчальної вибірки (рис. 2.5). При перенавчанні помилка на навчальній вибірці буде мінімальною, проте високою на тестовій.

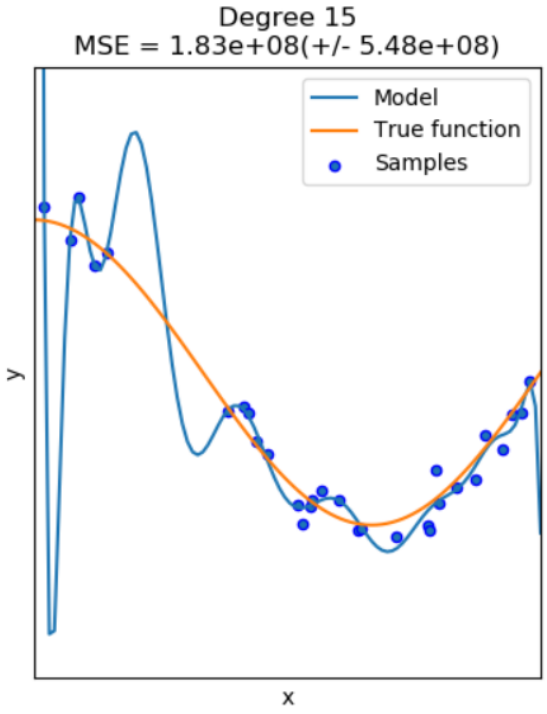

Рисунок 2.5 Занадто складна, перенавчена модель (overfitting) [27]

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

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

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

Найпопулярнішим методом являється є градієнтний спуск. Суть методу:

Повторювати доки помилка > ε або досягнута певна кількість ітерацій:

Одночасно змінювати усі ваги за наступною формулою:

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

де α – параметр швидкості навчання.

Або у матричній формі:

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

Проте даний метод має свої недоліки:

- необхідно підбирати α;
- шукає локальні мінімуми;
- при великій різниці між значеннями змінних навчання буде проходити повільніше (необхідно використовувати нормалізацію, або інші методи попередньої обробки).

Інший метод знаходження θ називається нормальне рівняння (Normal equation) й описується наступною формулою:

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

Даний метод має такі переваги:

- не потребує ітерацій;
- дає максимально можливу точність;
- швидкість не залежить від різниці між значеннями змінних.

Проте має й свої недоліки:

- не завжди можна знайти обернену матрицю, а отже даний метод не завжди може бути застосований;
- при достатньо великій кількості змінних працює повільно, оскільки складність алгоритму знаходження оберненої матриці $O(n^3)$.

2.1.3 Нейронні мережі

Штучна нейронна мережа – метод машинного навчання, який являє собою мережу зі штучних нейронів, які отримують сигнал на вхід, на основі якого змінюють свій внутрішній стан, й створюють вихідні значення (які залежать від внутрішнього стану, та входу) [28].

В основі штучних нейронів лежать їх біологічні аналоги. Саме тому ще на початку історії їх створення було розділення на 2 напрямки. Перший був сфокусований на моделюванні біологічних процесів, а інший на їх практичне використання у задачах [29].

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

1) на кожен вхід нейрона (x_j) подається сигнал (може подаватися від нейронів попередників i);

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

3) відбувається додавання отриманих значень;

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

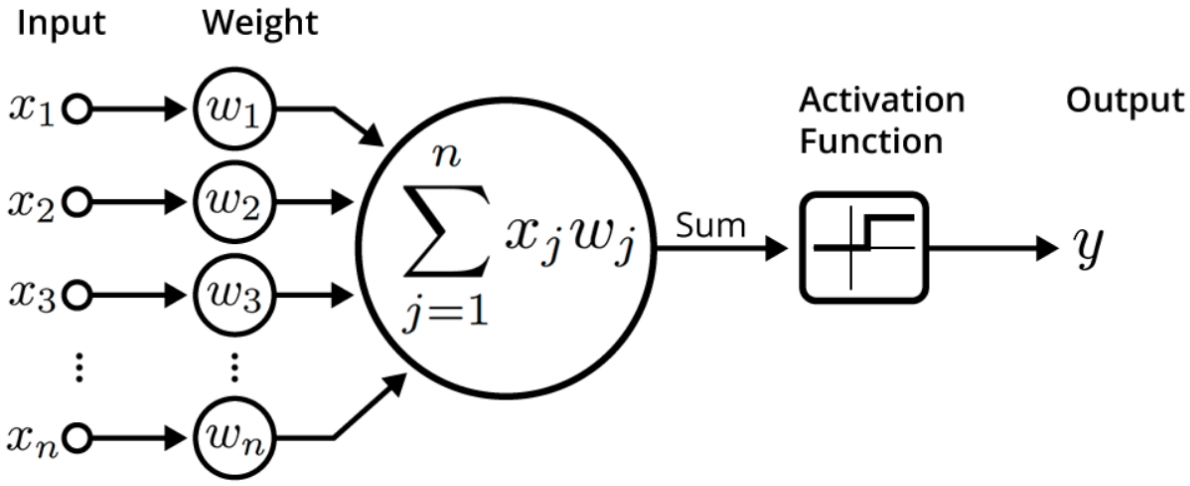

Рисунок 2.6 Ілюстрація штучного нейрона [30]

У якості функції активації обираються нелінійні функції, оскільки вони дозволяють вирішувати нетривіальні задачі. Так, наприклад, згідно з універсальною теоремою апроксимації (теорема Цибенко), нейронна мережа прямого поширення з одним прихованим шаром, з сигмоїдною функцією активації може апроксимувати будь-яку неперервну функцію багатьох змінних з довільною точністю [31]. Якщо використовувати лінійні функції – то нейронні мережі розв'язували б дуже обмежений клас задач, які можуть розв'язувати лінійні апроксимуючі функції. Функції активації представлені 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	$f(x) = tanh\ tanh\,(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$	(-1, 1)
Функція Гауса (дзвоноподібна)	$f(x) = e^{-x^2}$	(0, 1]
ReLU (напівлінійна)	$f(x) = \{0, x < 0\ x, x \geq 0$	[0, ∞)
Напівлінійна з насиченням	$f(x) = \{0, x \leq 0\ x, 0 < x < 1\ 1, x \geq 1$	[0, 1]

Порогова	$f(x) = \{0, x < 0\ 1, x \geq 0$	$\{0, 1\}$
Трикутна	$f(x) = \{1 - \lvert x\rvert, \lvert x\rvert \leq 1\ 0, \lvert x\rvert > 1$	$[0, 1]$

Розглянемо найпопулярніші функції активації більш детально.

Найпопулярнішою функцією активації являється *сигмоїда* (рис. 2.7). Вона є нелінійною, гладкою й монотонно зростаючою функцією.

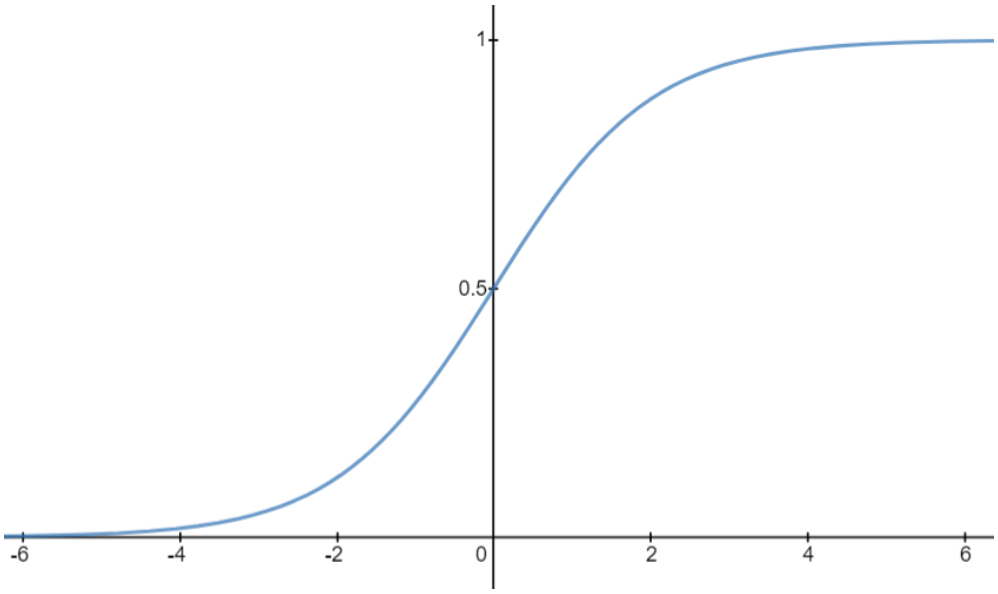

Рисунок 2.7 Графік сигмоїдної функції

Має наступні переваги:

- має область значень (0,1), що дає нормалізацію вихідного значення для кожного нейрона;
- має гладкий градієнт, який запобігає різким змінам (стрибкам) при підрахунку значень;
- у діапазоні x від -2 до 2 значення y швидко змінюється й має тенденцію до наближення до одної з асимптот, що дозволяє робити чіткі передбачення класу.

Проте має й недолік: при приближенні до асимптот сильно зменшується значення похідної, що негативно впливає на швидкість навчання.

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

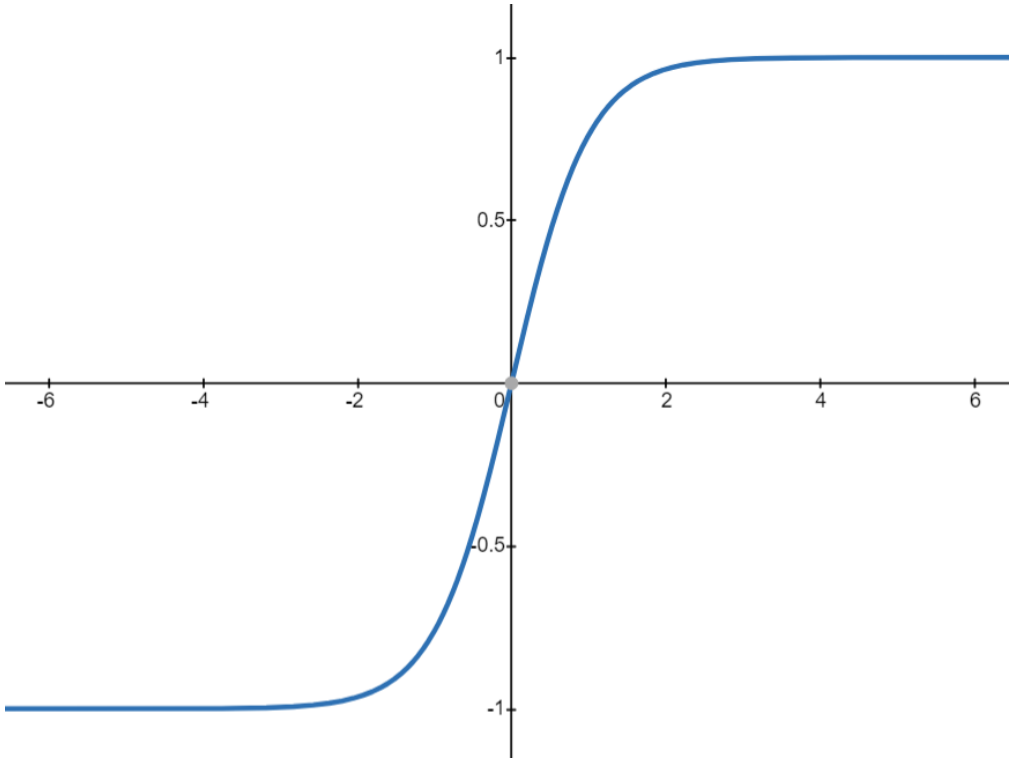

Рисунок 2.8 Графік гіперболічного тангенсу

Переваги:

- усі переваги сигмоїди, крім нормалізації;
- має негативні значення, що може бути корисно при роботі з ними;
- у порівнянні 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 являється нелінійною.

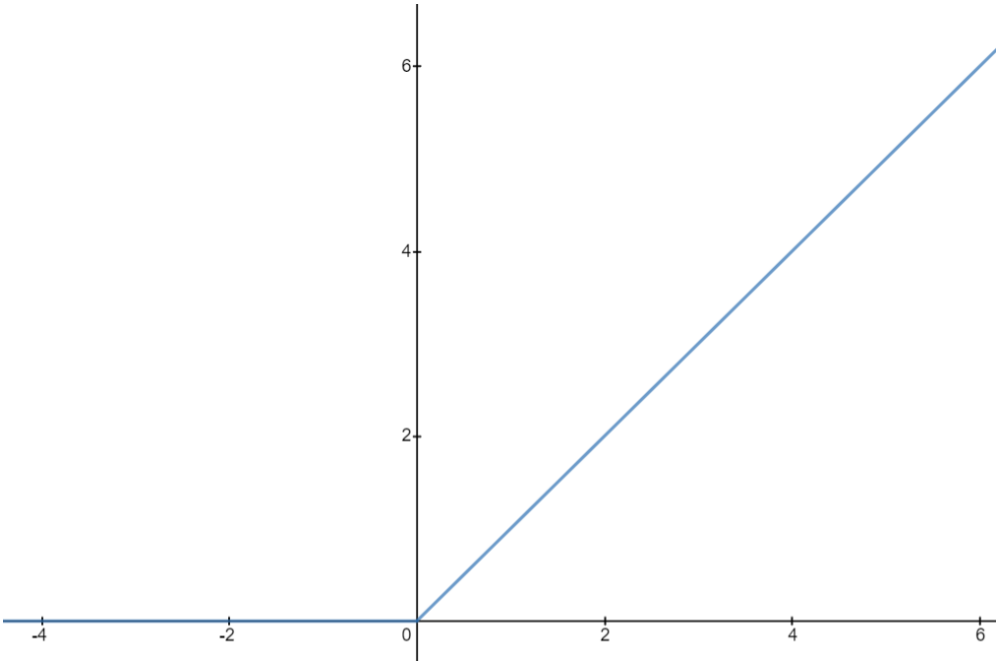

Рисунок 2.9 Графік ReLu

Переваги ReLu:

- швидко та просто розраховується похідна;
- спрощує нейронну мережу оскільки не усі нейрони будуть активовані, що позитивно впливає на швидкість навчання.

Однак ReLu має недолік: частина функції має похідну 0, у результаті чого деякі ваги не будуть змінюватися під час навчання.

Неможливо однозначно відповісти, яку функцію активації краще обрати. Зазвичай функції обираються для певної задачі на основі їх особливостей (наприклад, для задачі класифікації добре підходить сигмоїда, а гіперболічний тангенс гірше), або за їх подібність до тих функцій, які намагаються апроксимувати.

Нейронна мережа формується шляхом з’єднання виходу певних нейронів з входами інших, у результаті чого утворюється зважений орієнтований граф. Залежно від того як з’єднані нейрони формується архітектура мережі. На сьогодні існує багато різних архітектур, які представлені на рис. 2.10. Тип архітектури нейронної мережі обирається залежно від задачі та необхідних особливостей.

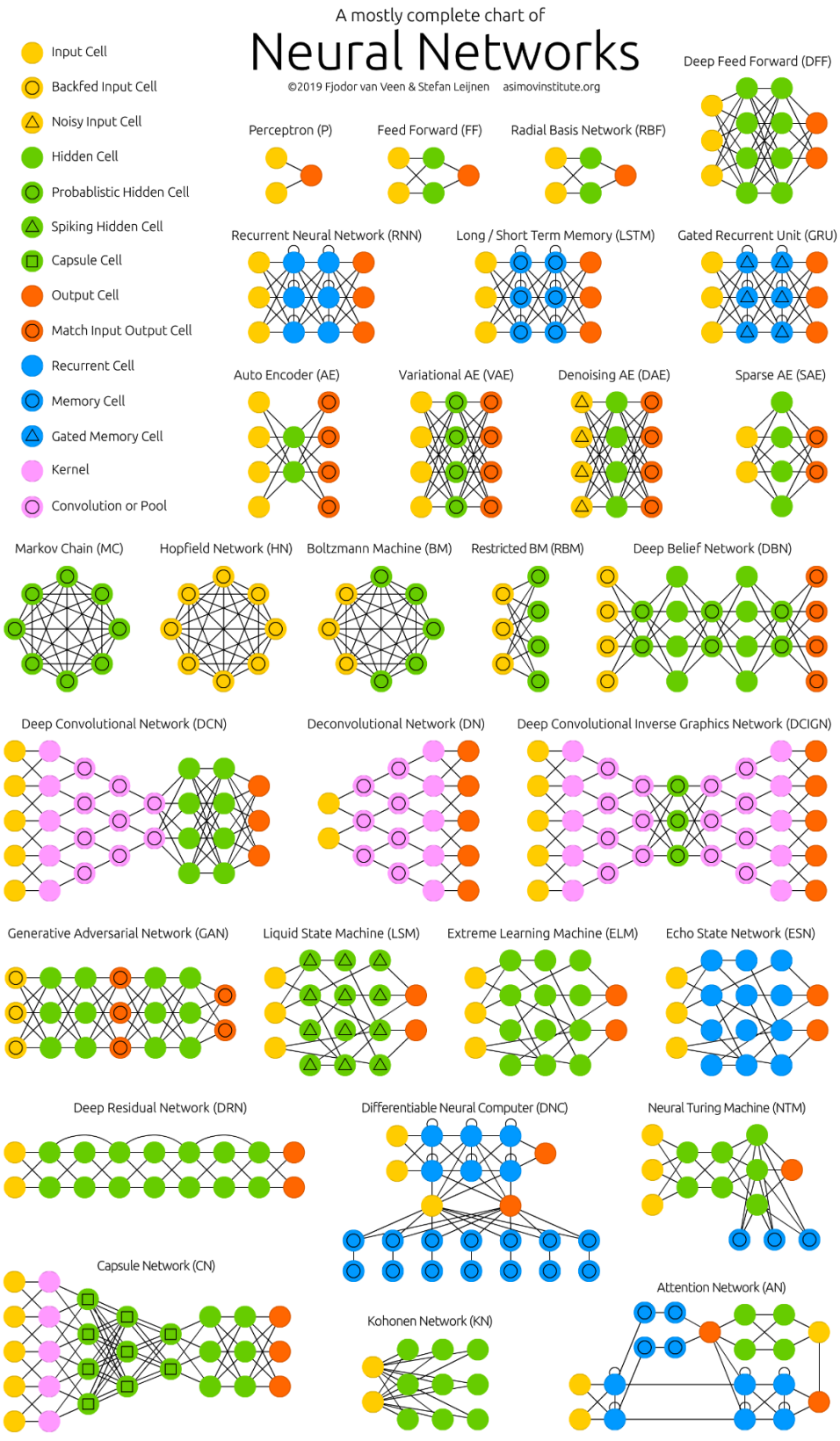

Рисунок 2.10 Класифікація архітектур нейронних мереж [28]

Окрім різниці у типах зв’язку, архітектури нейронних мереж можуть відрізнятися за кількістю прихованих шарів та кількістю нейронів. Не дивлячись

на універсальну теорему апроксимації, деякі функції дуже важко, або майже неможливо точно апроксимувати лише за допомогою одного прихованого шару.

Для точної роботи мережі оптимальна кількість шарів та нейронів зазвичай підбирається експериментальним шляхом. При замалій їх кількості мережа не зможе правильно віднайти залежність (underfitting), аналогією може слугувати апроксимація квадратичної залежності лінійною функцією. Якщо ж шарів або нейронів забагато – мережа буде перенавчена (overfitting), тобто вона буде «запам’ятовувати» випадки з навчальної вибірки, на яких помилка буде малою, проте не зможе давати точні результати на навчальній вибірці або реальних значеннях. Крім цього така мережа працює повільніше та довше навчається.

Однією з найперших та досі популярною архітектурою являється багатошаровий перцептрон (рис. 2.11), який є повнозв’язною мережею прямого поширення.

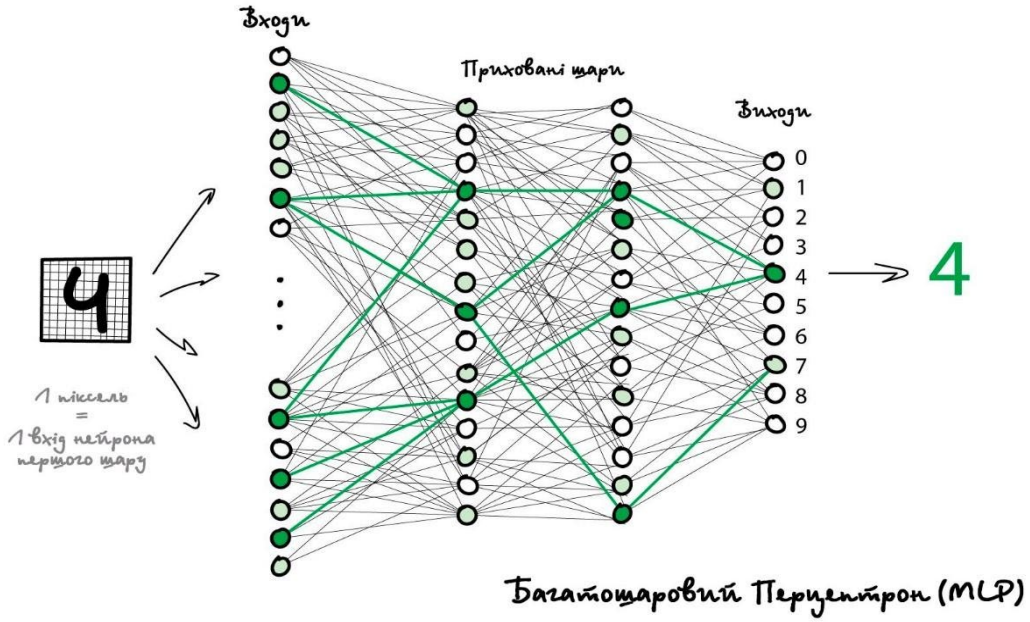

Рисунок 2.11 Приклад багатошарового перцептрону [23]

Нейронною мережею прямого поширення (Feed-Forward Neural Networks) називається мережа у якій сигнал проходить лише в одному напрямку – від

вхідного шару, через приховані шари, до вихідного шару на якому отримується результат у вигляді скаляра або вектора.

Повнозв’язною нейронною мережею є мережа у якій кожен нейрон зв’язаний з усіма нейронами з попереднього шару.

Переваги багатошарового перцептрону:

- проста архітектура;
- універсальність – може вирішувати різні задачі.

Недоліки:

- у більш складних задачах буде мати занадто багато параметрів, що робить його використання нераціональним;
- при великій кількості прихованих шарів відбувається «затухання» градієнта, у результаті чого навчання відбувається набагато довше.

Таким чином можна зробити висновок, що багатошаровий перцептрон добре підходить для простих задач класифікації або регресії, а для більш складних задач, наприклад розпізнавання зображень, необхідно обрати нейронну мережу з іншою архітектурою. Отже, така архітектура мережі підходить для вирішення нашої задачі.

Для коректної роботи нейронної мережі необхідно мати правильні ваги між усіма нейронами. Для отримання таких ваг використовуються алгоритми навчання. Залежно від архітектури мережі, алгоритми навчання можна класифікувати таким чином:

- навчання з вчителем;
- навчання без вчителя;
- змішане навчання;
- навчання з підкріпленням.

Перед початком навчання ваги необхідно ініціалізувати певними значеннями. Для цього існує декілька варіантів:

- використати ваги з попередньо навчених нейронних мереж;
- вибрати випадкові невеликі значення.

Найпопулярніший алгоритм навчання – метод зворотного поширення помилки. Основна суть методу в тому, що після отримання результату він порівнюється з реальним значенням. Після чого розраховується помилка й у зворотному напрямку отримується внесок кожного нейрону у неї. На основі таких внесків й коригуються кожні ваги окремо.

Алгоритм методу:

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

Для усіх прикладів з навчальної вибірки:

1. подати вхідні значення та отримати вихідні значення з кожного нейрону (a_i);
2. для кожного вихідного нейрону розраховується помилка $\delta_k = f'(a_k) * (y_k - a_k)$, де f' – похідна від функції активації, y_k – реальне значення, $k \in$ кількість вихідних нейронів;
3. для кожного наступного шару i ($i \in$ кількість шарів $- 1$): Для кожного нейрону j ($j \in$ кількість нейронів у i) розраховується

4. $\delta_j = f'(a_j) * \sum_{k \in \text{нейрони з } i+1} \delta_k w_{j,k}$.

5. для кожного w_{ij}:

6. $\Delta w_{ij} = \alpha * \delta_j * a_i; w_{ij} := w_{ij} + \Delta w_{ij}$,

7. де α – параметр швидкості навчання;
8. повернути значення w_{ij}.

Таким чином, створення нейронної мережі можна звести до таких етапів:

1. попередня обробка даних(за необхідністю);
2. створення навчальної та тестової вибірки;
3. вибір архітектури мережі;
4. вибір алгоритму навчання;
5. вибір параметрів навчання;
6. навчання мережі;
7. перевірка точності мережі;
8. при необхідності скоригувати параметри мережі або провести повторне навчання.

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 – ієрархічні деревоподібні структури, які складаються з вирішальних правил «якщо–то» й дозволяють виконувати класифікацію об’єктів (рис. 2.12) [33].

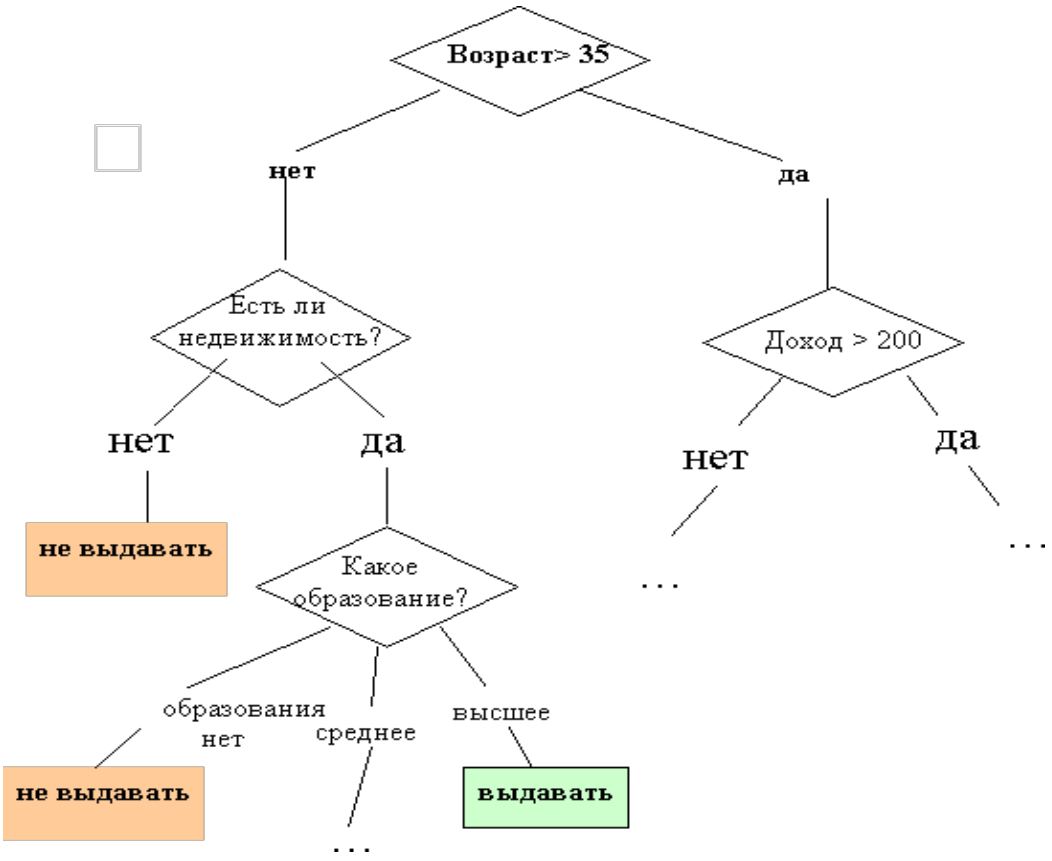

Рисунок 2.12 Приклад дерева рішень [34]

Дерева рішень мають два типи об’єктів – вузли (node) та листя (leaf). У вузлах містяться правила, за допомогою яких виконується перевірка ознак й відбувається розбиття множини об’єктів на підмножини. Листя – це кінцеві вузли дерева, у яких містяться підмножини асоційовані з конкретними класами. Основна відмінність листя від вузлів у тому, що там не проводиться перевірка й відсутнє подальше розгалуження [33].

Як й інші моделі, дерева рішень будуються на основі навчальної вибірки. У процесі побудови дерева формуються правила рішень і для кожного такого правила створюється вузол. Для кожного вузла необхідно обрати ознаку (атрибут), по якій буде проводитися перевірка правила. Обирати ознаки необхідно таким чином, щоб забезпечити найкраще розбиття у вузлі. Найкращим розбиттям вважається те, яке дозволяє класифікувати найбільшу кількість об’єктів і створює максимально «чисті» підмножини [33].

Ансамблеві методи – це методи, які використовують одночасно декілька моделей з метою отримання більшої точності, порівняно з кожною моделлю окремо [35]. Залежно від виду ансамблю, моделі можуть базуватися як на однакових, так і на різних методах навчання. Моделі, які використовуються в ансамблі повинні відрізнятися одна від одної. Крім того, зазвичай обирають методи навчання, які «нестабільні», наприклад, чуттєві до викидів або аномалій.

Ефективність ансамблевих методів можна розглянути на основі теореми Кондорсе «Про журі присяжних» [36]. Якщо кожен член журі присяжних має незалежну думку, і якщо ймовірність правильного рішення члена журі більше за 0.5, тоді ймовірність правильного рішення присяжних зростає зі збільшенням кількості членів журі, і наближається до одиниці. Математично це можна записати таким чином:

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

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

Беггінг (Bagging, від Bootstrap aggregation) – один з видів ансамблів, який базується на статистичному методі бутстрепу [37].

Метод бутстрепу полягає у наступному. Нехай існує вибірка X розміром N. З вибірки випадково вибирається N об’єктів з поверненням й створюється підвибірка X_1. Таким чином, на кожній спробі у кожного об’єкта ймовірність щоразу бути вибраним складає $\frac{1}{N}$. Об’єкти у підвибірці можуть повторюватися. Повторюємо процедуру разів M й генеруємо підвибірки $X_1, \dots, X_M$.

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 об’єктів. На кожному кроці кожен об’єкт потрапляє у

підвибірку з ймовірністю $\frac{1}{N}$. Отже, ймовірність того що об'єкт не попаде у підвибірку складає $(1-\frac{1}{N})^N$. При $N \to \infty$ отримуємо ймовірність $\frac{1}{e} \approx 37\%$. А ймовірність кожного об'єкту попасти у підвибірку складає $1-\frac{1}{e} \approx 63\%$.

Суть беггінгу виглядає наступним чином (рис. 2.13). На основі вибірки X генеруються підвибірки $X_1, \dots, X_M$. На кожній підвибірці тренується модель $a_i(x)$. Кінцева модель буде усереднювати відповіді кожної моделі: $a(x) = \frac{1}{M}\sum_{i=1}^{M} a_i(x)$, або у випадку класифікації проводити голосування.

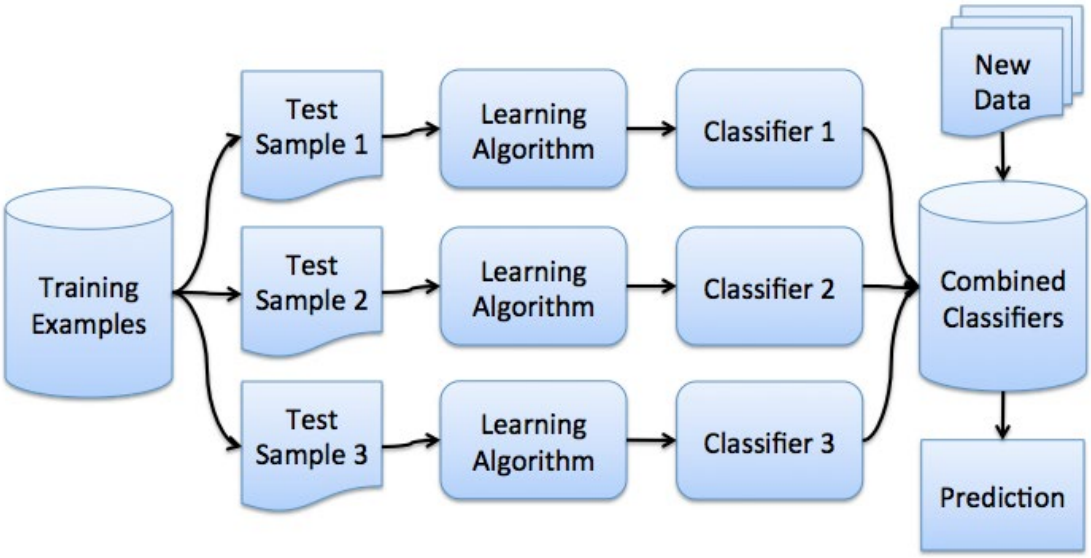

Рисунок 2.13 Приклад ансамблевого методу на основі беггінгу [38]

Ефективність беггінгу забезпечується завдяки тому, що моделі, які пройшли навчання на різних підвибірках, виходять досить різними і їх помилки взаємно компенсуються. Крім того, викиди у даних можуть не потрапляти у деякі підвибірки. Беггінг ефективний на малих вибірках, коли втрата навіть невеликої частини об'єктів призводить до отримання різних моделей. У разі великих вибірок можна генерувати підвибірки меншого розміру.

Крім беггінгу, випадковий ліс застосовує метод випадкових підпросторів [39]. У даному методі моделі навчаються на підмножинах ознак, які обираються випадковим чином. Даний метод дозволяє знизити кореляцію між деревами та

знижує ймовірність перенавчання. Алгоритм побудови ансамблів на основі методу випадкових підпросторів виглядає таким чином:

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

У випадковому лісі для задач регресії рекомендується обирати $d = \frac{D}{3}$, а для задач класифікації $d = \sqrt{D}$. Проте у різних задач оптимальне значення 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 $n = 1, \dots, N$, де N – необхідна кількість дерев у ансамблі:

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

- по заданому критерію обирається краща ознака, по якій проводиться розбиття у дереві до вичерпання підвибірки;
- при кожному розбитті обирається d випадкових ознак серед 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 $a(x) = \frac{1}{N}\sum_{i=1}^{N} a_i(x)$.

2.2 Аналіз основних видів вхідних даних та їх вибір

2.2.1 Вибір вхідних даних

В якості даних для сільського господарства зазвичай використовуються кількісні (чисельні) дані, такі як кількість добрив, опадів, температура тощо. В останні десятиліття при прогнозуванні врожайності сільськогосподарських культур, зокрема посівів зернових, все більше застосування, поряд з наземної інформацією, отримують дані спостереження Землі з космосу або дистанційного зондування Землі, наприклад безпілотними апаратами. Спостереження з космосу за допомогою супутників з впровадженням чутливих датчиків за параметрами та характеристиками поверхні Землі проводиться ще з кінця минулого століття. Отримані дані відкривають нові можливості моніторингу і контролю врожайності сільськогосподарських культур. Крім традиційного аерофотографування використовуються й більш складні дані, такі як NDVI (Normalized Difference Vegetation Index), VHI (Vegetation Health Index) тощо.

При створенні моделей прогнозування такі комплексні показники надають значну перевагу, оскільки за рахунок використання малої кількості показників розмірність вхідних даних невелика, що спрощує моделі та позитивно впливає на їх точність при вибірці невеликих розмірів. Тому було прийнято рішення використовувати NDVI та VHI у якості ознак разом з кількістю внесених добрив.

2.2.2 Індекс NDVI

Нормалізований диференційний вегетаційний індекс (Normalized Difference Vegetation Index, NDVI) – показник кількості фотосинтетичної активної біомаси [41].

NDVI часто використовується по всьому світу для моніторингу засухи, моніторингу та прогнозування сільськогосподарського виробництва, надання

допомоги в прогнозуванні небезпечних зон пожеж і карт наступу пустелі. NDVI краще для глобального моніторингу рослинності, бо дає змогу компенсувати зміну умов освітлення, нахил поверхні, експозицію та інші зовнішні чинники [42].

Розраховується NDVI за формулою:

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

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

Для відображення значень індексу використовується неперервний градієнт або дискретна шкала. Значення індексу NDVI лежить в межах від -1 до 1. Значення індексу для рослинності зазвичай лежить в діапазоні від 0,2 до 0,95 (рис. 2.14). Чим краще розвинута рослинність під час вегетації, тим вище значення NDVI.

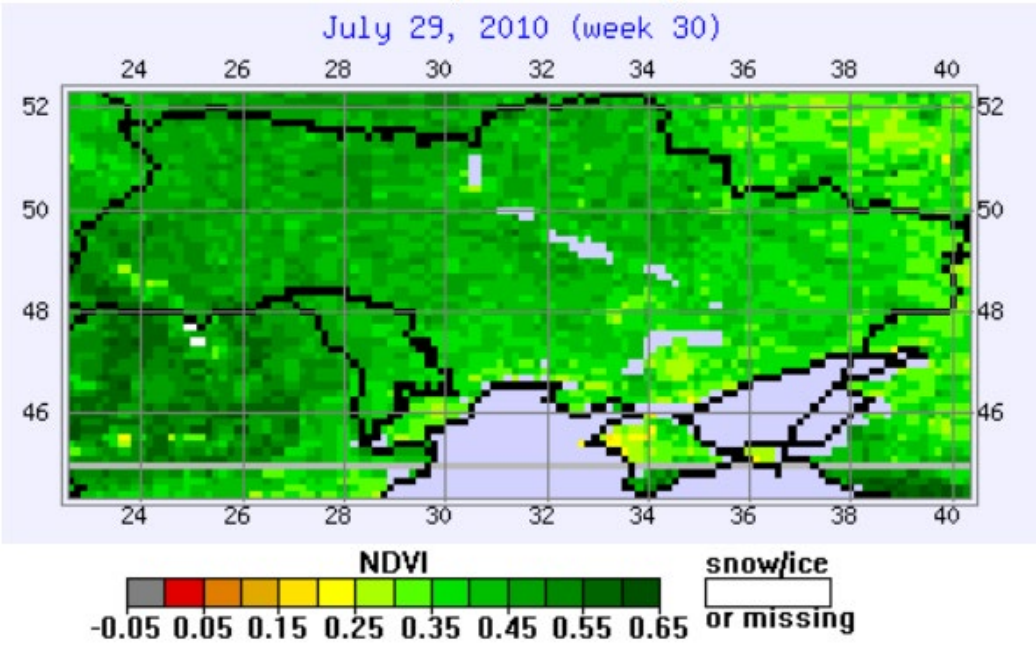

Рисунок 2.14 Приклад графічного відображення NDVI з дискретною шкалою [43]

Таким чином, NDVI – це індекс, за яким можна судити про розвиток зеленої маси рослин під час вегетації. Завдяки тому, що відображувальна здатність

нерослинних об’єктів не залежить від пори року – їх індекс NDVI має фіксоване значення, яке нижче, порівняно з рослинами (рис. 2.15).

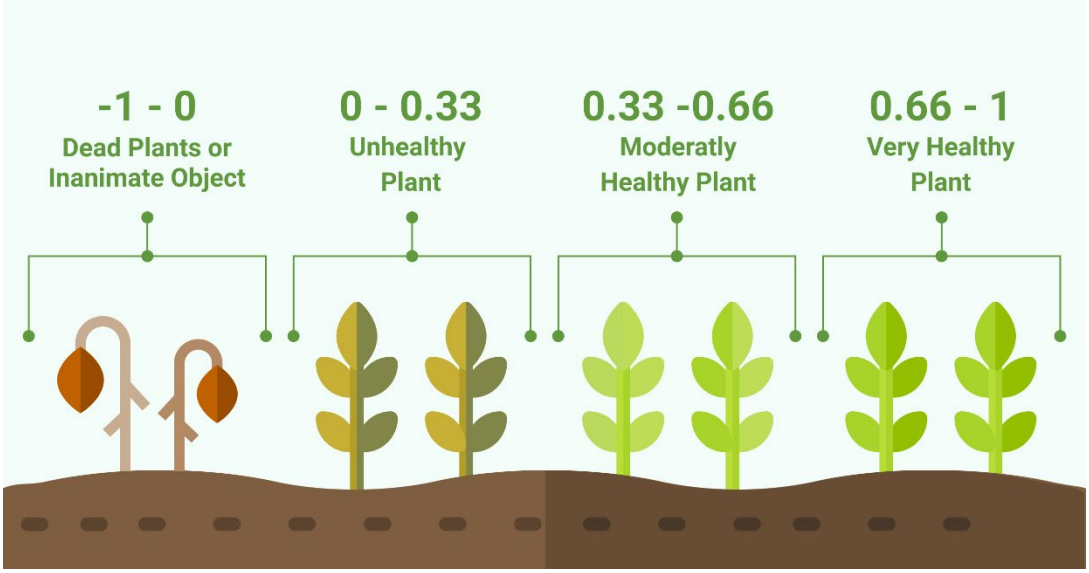

Рисунок 2.15 Залежність NDVI від стану здоров’я рослин [44]

Протягом вегетації індекс NDVI зростає, досягає свого піку близько 0,80-0,85 (у зернових – це момент колосіння) і потім починає знижуватися (рис. 2.16). Зниження індексу в кінці вегетації відображає процес дозрівання культур. Тому, наприклад, для декількох полів зернових культур за індексом NDVI можна визначити найбільш оптимальний порядок збирання полів – чим нижчий індекс, тим сухіше зерно.

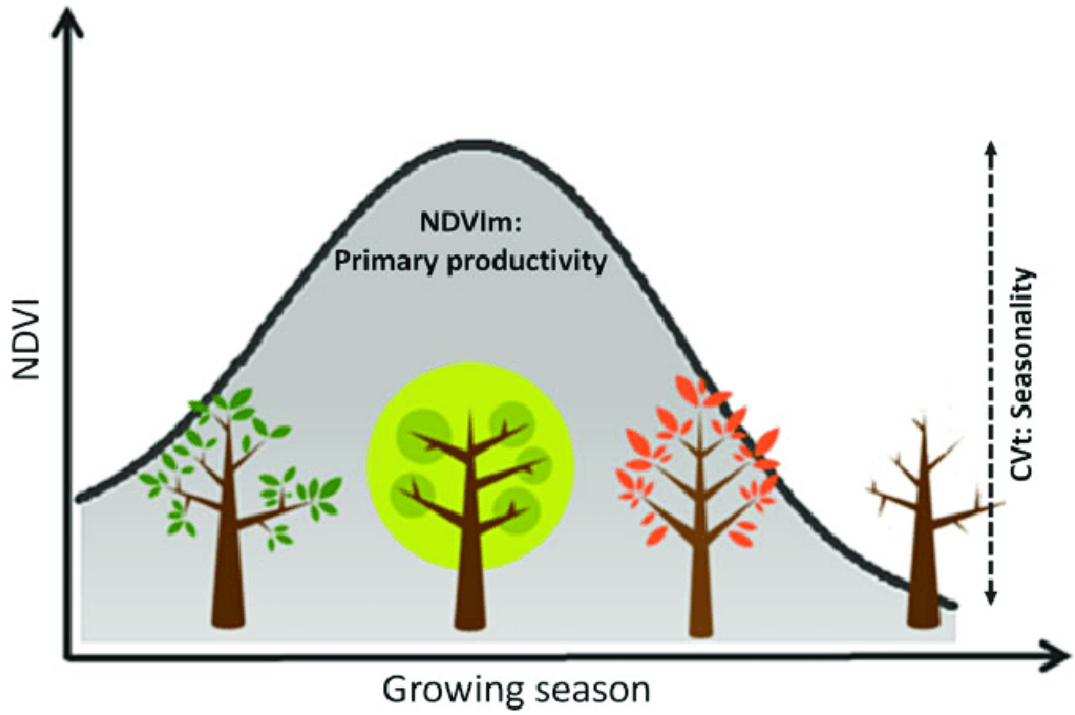

Рисунок 2.16 Вплив фази розвитку рослин на індекс NDVI [45]

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

2.2.3 Індекс VHI

Вегетаційний індекс здоров'я (Vegetation Health Index, VHI) – показник, який характеризує здоров'я рослинності, припускаючи що стресові умови рослин пов'язані з нижчим за нормальний рівень NDVI та вищою за нормальну температуру [46, 47].

Розраховується VHI за формулою:

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

VHI широко застосовується як індекс посухи на основі даних дистанційного зондування й розраховується як зважена сума двох компонентів: індекс стану рослинності (Vegetation Condition Index, VCI) та індекс теплового

стану (Thermal Condition Index, TCI). Використання комбінації індексів TCI та VCI надає покращене представлення рівня посухи [48].

VCI характеризує рівень вологості і зазвичай базується на даних з видимого та близького інфрачервоного електромагнітного спектру. Індекс має високу точність у якості оцінки посухи та базується на її впливі на тривалість та інтенсивність вегетаційного періоду [48]. VCI може використовуватися з іншими показниками для передбачення стану рослинності. Розраховується індекс за формулою:

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

де $NDVI'$ – середнє значення NDVI, $NDVI_{min}$ – найменше значення індексу, $NDVI_{max}$ – 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:

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

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

Оптимальні вагові коефіцієнти які застосовуються до VCI та TCI при розрахунку VHI зазвичай невідомі, тому припускається, що вагові коефіцієнти мають значення 0,5 для кожного індексу, таким чином α = 0,5. Недавні дослідження виявили, що можливо покращити точність вагових коефіцієнтів при

порівнянні результуючого VHI та несупутникових індикаторів посухи, таких як SPEI [49].

Аналіз даних виявляє кореляцію між TCI та VCI (LST-NDVI). Сильна кореляція дозволяє використовувати дані індекси як базис при розрахунку VHI. Проте при слабкій кореляції використання TCI та VCI недоречно й тому не можна розрахувати VHI, отже даний індекс не може використовуватися для деяких ділянок у якості оцінки посухи [50].

Для відображення значень індексу використовується неперервний градієнт або дискретна шкала. Значення індексу VHI лежить в межах від 0 до 100 (рис. 2.17). Чим вище значення індексу, тим здоровіша рослинність.

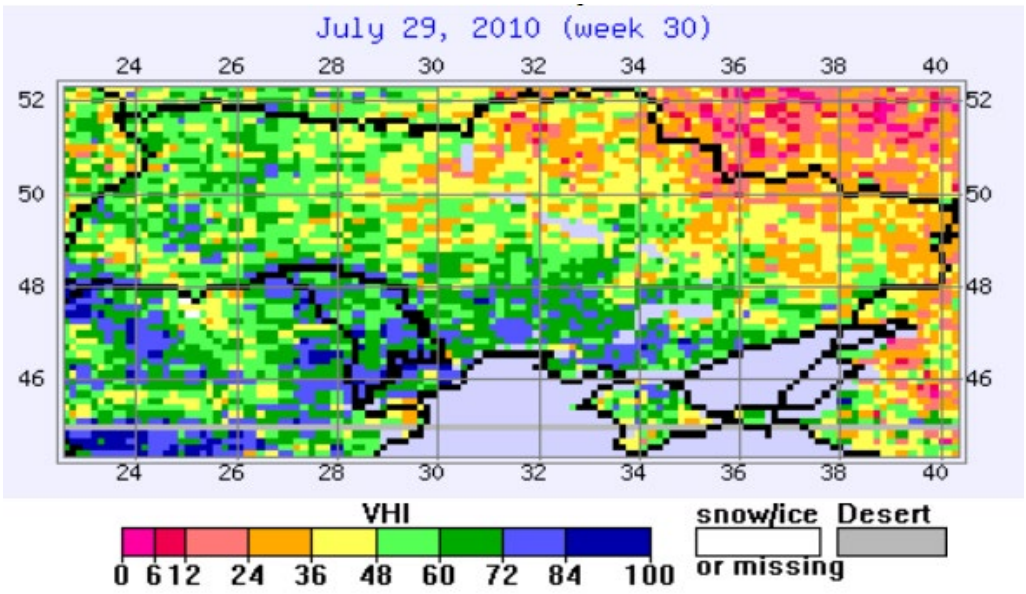

Рисунок 2.17 Приклад графічного відображення VHI з дискретною шкалою [43]

Отже, VHI може використовуватися у якості вхідних даних 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

Перед початком побудови моделей для прогнозування врожайності, корисним буде провести аналіз даних, які використовуються при створенні вибірок. Результати аналізу можуть надати підказку й допомогу у виборі необхідної архітектури.

Як було сказано раніше, у якості вхідних даних (незалежних змінних) використовуються індекси NDVI, VHI та кількість внесених мінеральних добрив (кг/га). У якості вихідних даних (залежної змінної) використовується врожайність певної сільськогосподарської культури (ц/га). У якості сільськогосподарських культур було обрано зернові культури, соняшник та цукровий буряк. Дані бралися по Україні за період 1992-2017 рр.

У зв’язку з невеликою розмірністю вхідних даних, першим кроком в аналізі буде їх візуалізація (рис. 3.1-3.3). Для візуалізації будуть використовуватися матриці графіків. На діагоналі матриці розташовані гістограми розподілу ознак, а інші елементи матриці містять діаграми розсіювання відповідних пар ознак. Для виконання візуалізації використовується бібліотека *seaborn* для мови *Python*, яка базується на бібліотеці *matplotlib*, та бібліотека *pandas* для взаємодії з даними.

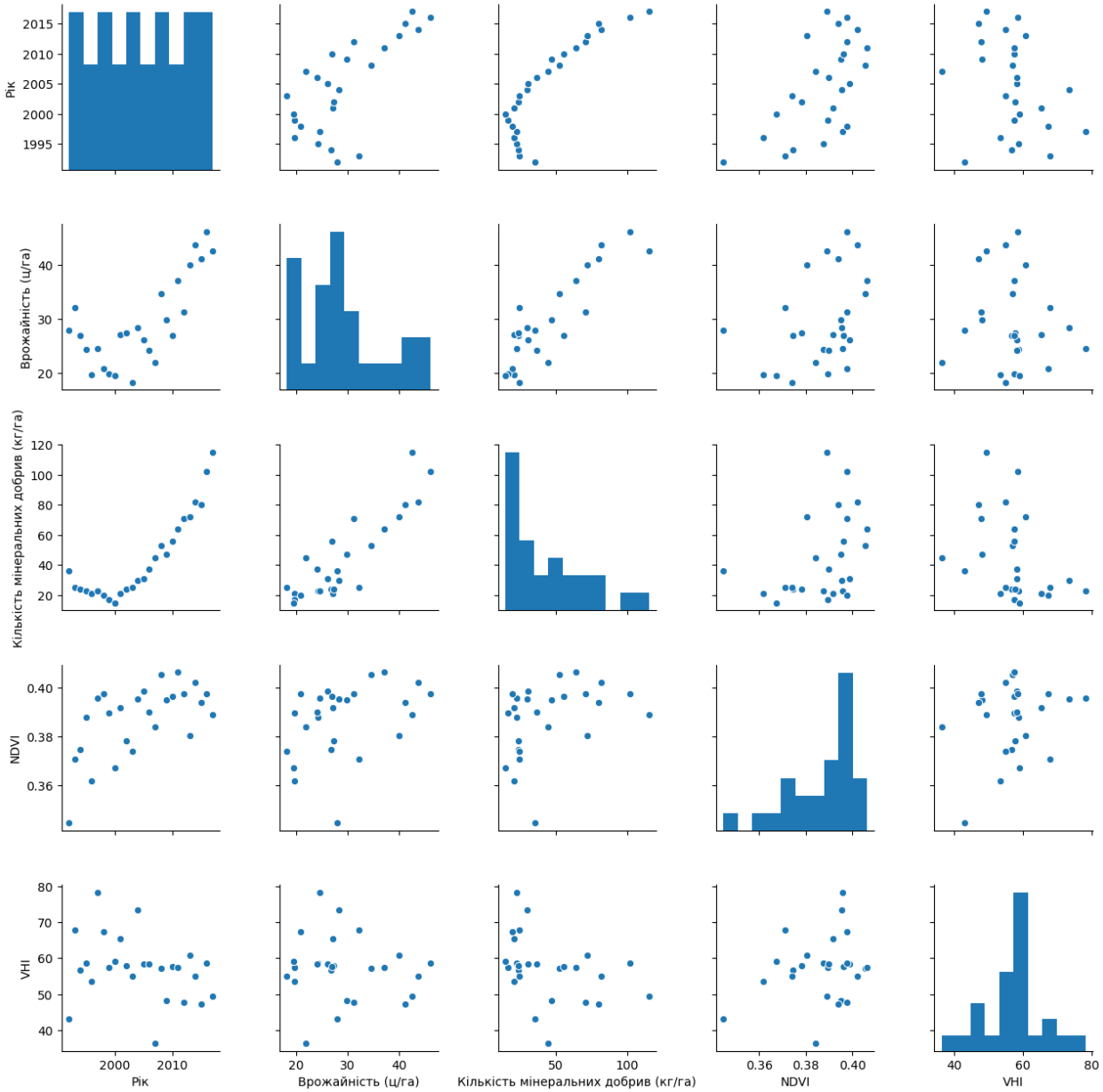

Рисунок 3.1 Матриця графіків для зернових культур

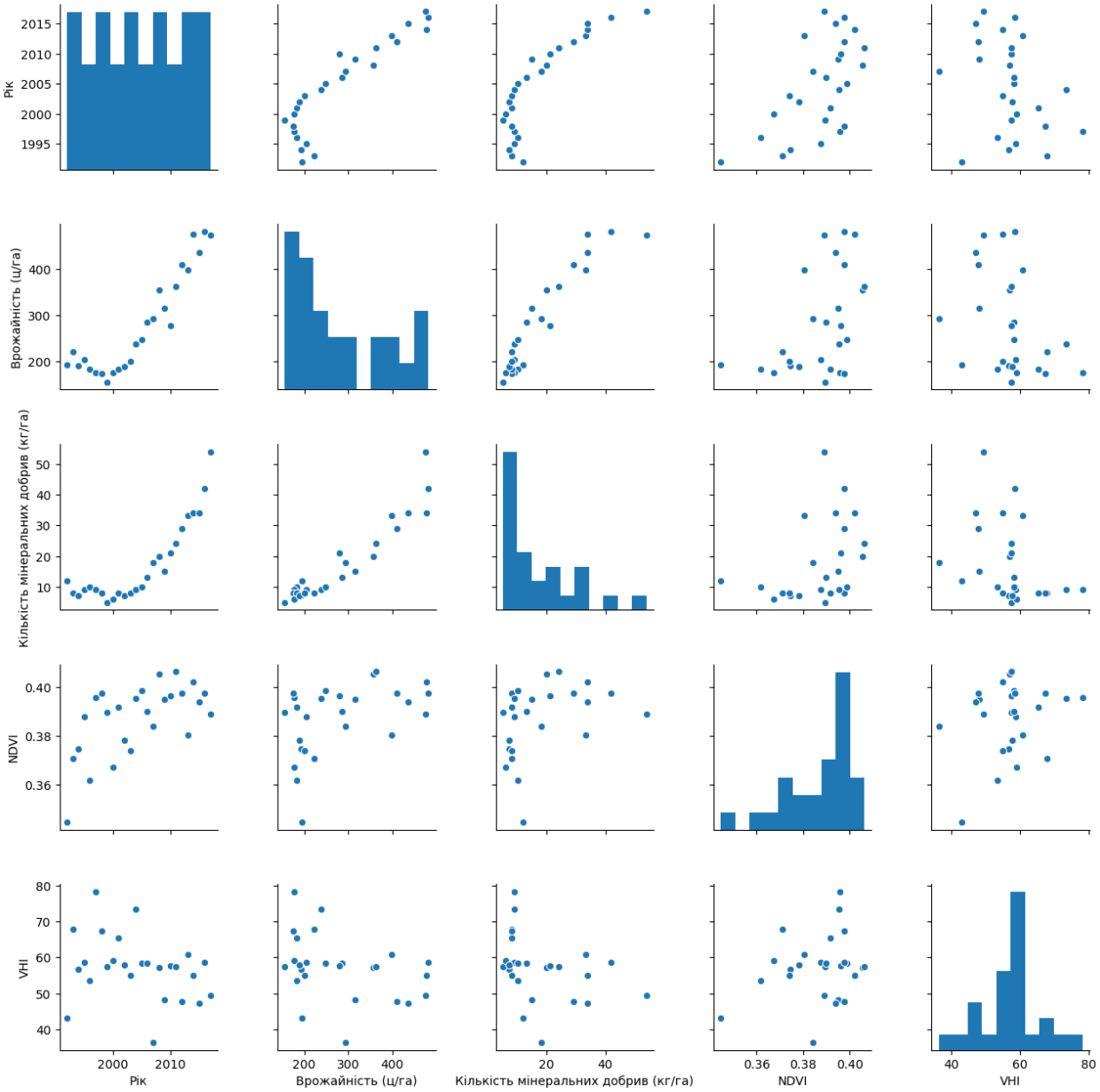

Рисунок 3.2 Матриця графіків для цукрового буряку

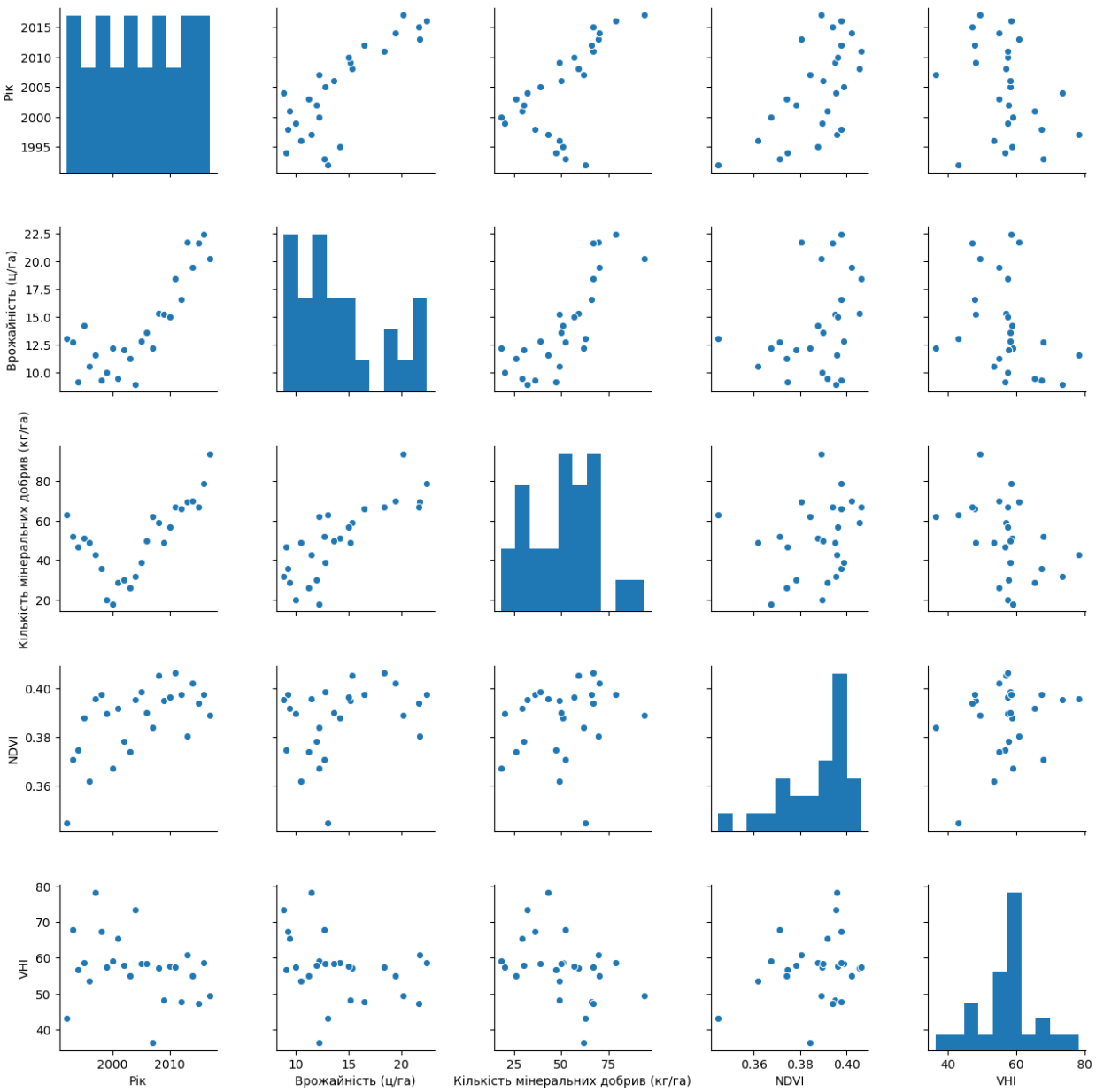

Рисунок 3.3 Матриця графіків для соняшнику

При візуальному аналізі отриманих результатів можна відмітити доволі чітку, подібну на лінійну, залежність між врожайністю та кількістю внесених добрив для всіх сільськогосподарських культур. Щодо інших важливих показників, не можна зробити однозначних висновків про залежності на основі візуального аналізу.

Цікаво відмітити, що існує дуже чітка нелінійна залежність між парами ознак «рік» – «кількість мінеральних добрив» та «рік» – «врожайність». Подібну залежність вже знаходили іноземні дослідники, а саме позитивний лінійний

тренд, за останні десятиліття, у врожайності озимої пшениці, який пов’язаний з покращенням сільськогосподарських технологій [51]. Можна припустити, що нелінійний характер знайденої залежності пов’язаний з економічними проблемами України на початку її незалежності.

Для точнішого аналізу використаємо коефіцієнт кореляції Пірсона, який показує міру лінійної залежності, між залежною та незалежними змінними. Для розрахунку кореляції було використано бібліотеку *numpy*. Отримані результати занесені у табл. 3.1.

Таблиця 3.1 Результати кореляційного аналізу

	Кількість мінеральних добрив	NDVI	VHI
Зернові культури	0,87	0,38	-0,11
Цукровий буряк	0,81	0,32	-0,32
Соняшник	0,94	0,46	-0,36

Кореляційний аналіз підтвердив сильну лінійну залежність між врожайністю та кількістю внесених добрив. Інші ж показники мають слабку кореляцію, а VHI навіть негативну. Проте слабка кореляція може означати, що існує залежність нелінійного типу.

3.1.2 Вибір архітектури

Для вибору оптимальної архітектури будуть протестовані різні її варіанти. Той варіант архітектури, який отримає найкращу оцінку буде обраний для подальших експериментів.

Для тестування моделей було обрано метод перехресної перевірки або кросс-валідація (cross-validation). Суть методу полягає у розбитті вибірки на K

блоків однакового розміру. Один з блоків використовується як тестова вибірка, а інші $K-1$ утворять навчальну вибірку. На утвореній навчальній вибірці відбудеться навчання моделі, а для оцінки точності використовується тестова вибірка. Процес повторюється K разів, причому кожен блок буде використовуватися як тестова вибірка лише один раз. У результаті буде отримано K оцінок, а результуюча оцінка розраховується шляхом усереднення. Для кожної перевірки буде використовуватися $K = 5$.

У якості оцінки буде використано середньоквадратичну помилку, яка розраховується за формулою $MSE = \frac{1}{n}\sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)})^2$, де n – кількість випадків, $y^{(i)}$ – реальне значення у i-му випадку, $h_\theta(x^{(i)})$ – прогнозоване значення для i-го випадку. Таким чином, чим нижча оцінка, тим кращий результат.

В якості програмного забезпечення використовується бібліотека *scikit-learn*, яка містить усі методи машинного навчання описані раніше та необхідні інструменти для проведення тестування. Додатково використовується бібліотека *pandas* для роботи з даними.

Розпочнемо з поліноміальної регресії. Аналіз даних показав, що не між усіма незалежними та залежною змінною є чіткий лінійний зв'язок, тому звичайна лінійна регресія може бути не самим оптимальним варіантом. Для тестування будуть використовуватися поліноміальні регресії різного степеня, починаючи з першого (лінійного варіанту). Результати занесені в табл. 3.2.

Таблиця 3.2 Результат тестування поліноміальної регресії різних степенів

Степінь поліному	MSE		
	Зернові культури	Цукровий буряк	Соняшник
1	16,9	1680	15,2
2	$1{,}3 * 10^5$	$4{,}6 * 10^9$	$9{,}3 * 10^4$

3	$1,3 * 10^{13}$	$7,7 * 10^{10}$	$2,7 * 10^{6}$

На основі отриманих результатів можна зробити висновок, що для моделювання врожайності кожної сільськогосподарської культури оптимальні саме лінійні моделі (поліном першого степеня). При зростанні степеня поліному помилка збільшується на кілька порядків, що свідчить про перенавчання моделі.

Наступним методом для тестування буде нейронна мережа. При виборі архітектури, необхідно вдало підібрати кількість нейронів у прихованому шарі, аби їх було достатньо для пошуку залежності, проте ненабагато, щоб не відбулось перенавчання. У якості функції активації буде використовуватися ReLU, оскільки попереднє тестування показало, що при використанні сигмоїдальної функції мережа навчається повільніше й дає вищу помилку. Для зернових культур та соняшнику максимальна кількість ітерацій буде 10000, а для цукрового 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	Цукровий буряк	Соняшник
3	72,9	7125	18,6
5	16,6	2791	9,6
7	20,8	2403	9,8
9	20,3	1944	10,1
12	21,7	2834	10,6

Аналізуючи отримані результати, можна помітити загальну тенденцію: при малій кількості нейронів помилка висока. Зі збільшенням їх кількості, помилка

зменшується до певного рівня, а потім знову починає зростати. Таким чином, оптимальна кількість нейронів легко визначити по точці мінімуму помилки. Отже, оптимальні архітектури нейронних мереж для кожної сільськогосподарської культури виглядають таким чином:

- Зернові культури – 3/5/1;
- Цукровий буряк – 3/9/1;
- Соняшник – 3/5/1.

Останнім методом для тестування є випадковий ліс. Перевага випадкового лісу у тому, що при занадто великій кількості дерев рішень не відбудеться перенавчання. Проте якщо дерев замало – точність моделі буде недостатньою, а при збільшенні їх кількості – модель буде складнішою, потребуватиме більшої кількості пам’яті й працюватиме повільніше. Отже, необхідно оптимально підібрати кількість дерев у «лісі». Результати тестування занесені до табл. 3.4.

Таблиця 3.4 Результати тестування випадкового лісу з різною кількістю дерев рішень

Кількість дерев	MSE		
	Зернові культури	Цукровий буряк	Соняшник
10	59,5	5251	10,8
25	57,2	5175	9,9
50	56,7	4856	9,6
100	56,6	4808	9,6

Результати підтвердили те, що при збільшенні кількості дерев рішень зростає точність моделі, проте поступово швидкість зростання точності зменшується. Різниця у результатах між 50 та 100 деревами для кожної культури доволі незначна, що вказує на недоцільність подальшого збільшення кількості дерев. Отже, для кожної моделі на основі випадкового лісу буде використовуватися 100 дерев рішень.

3.2 Опис програмного додатку

3.2.1 Розробка програмного додатку

Для розробки програми спочатку необхідно створити та навчити моделі на основі різних методів машинного навчання. У якості цих методів використовуються:

- лінійна регресія;
- нейронна мережа;
- random forest.

Для кожного типу сільськогосподарської культури необхідно створювати окрему модель. Оскільки у роботі використовується 3 типи культур, то необхідно створити 9 різних моделей.

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

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;
- Цукровий буряк: архітектура 3/9/1, максимальна кількість епох – 100000, параметр швидкості навчання – 0,001, функція активації – ReLU;
- Соняшник: 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 (незалежних змінних) використовуються наступні: кількість внесених мінеральних добрив (кг/га), індекс NDVI, індекс VHI. В якості вихідних даних (залежної змінної) використовується врожайність культури

(ц/га). Дані було отримано зі звітів Державної служби статистики України [52] та супутникових даних NOAA STAR [43] (дані по Україні узяті за 1992-2017 роки).

Для створення моделей використовується бібліотека *Scikit-learn*, оскільки вона містить реалізації усіх необхідних методів, а для роботи з даними використовувалася бібліотека *pandas*. Приклад коду для завантаження даних, створення та навчання моделі наведений 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.

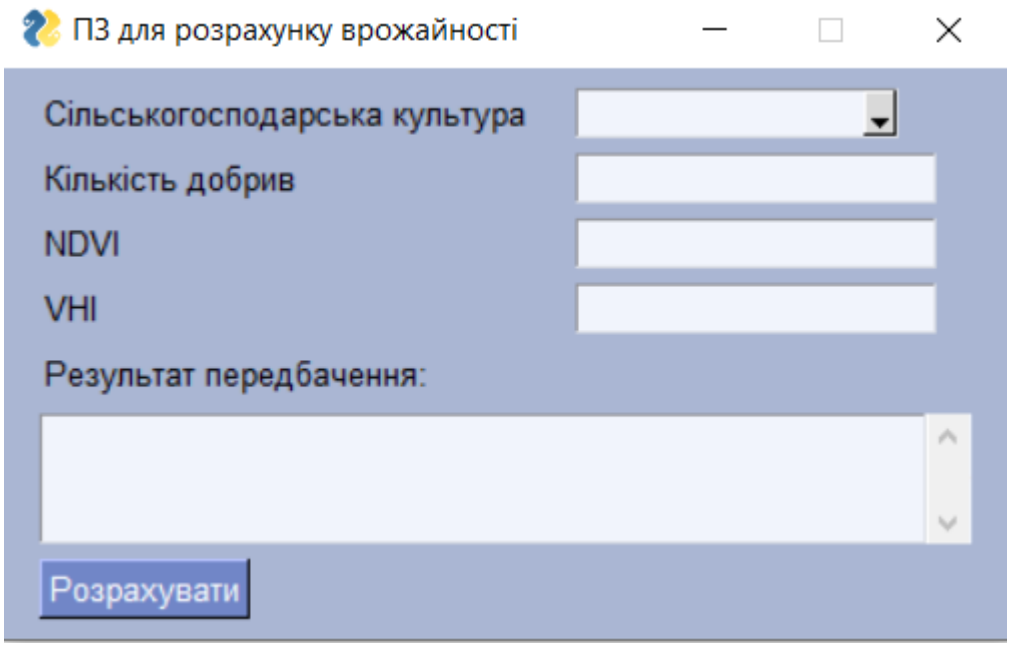

Рисунок 3.7 Інтерфейс ПЗ при запуску

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

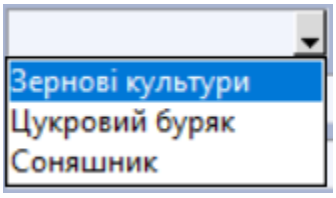

Рисунок 3.8 Випадаюче меню з переліком культур

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).

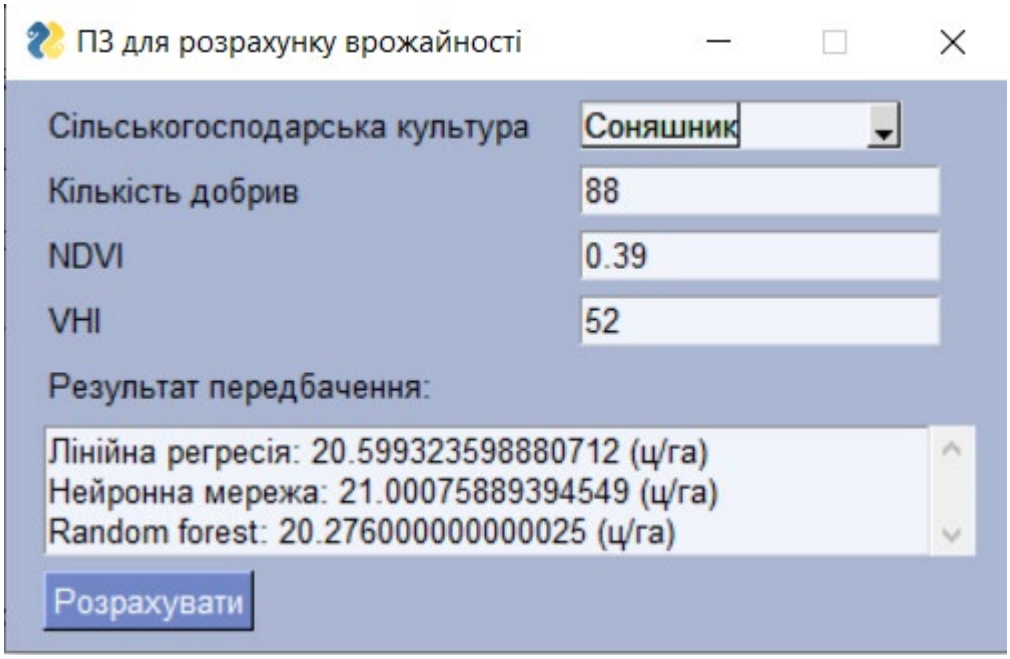

Рисунок 3.9 Приклад роботи ПЗ

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).

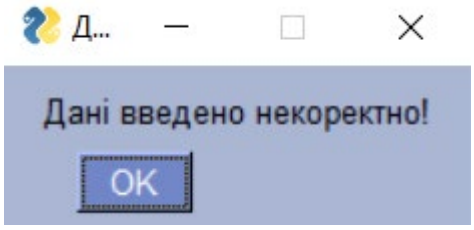

Рисунок 3.10 Вікно з попередженням про неправильність введених даних

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

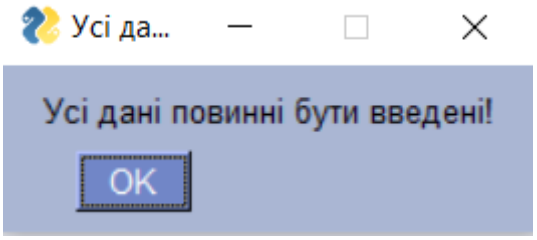

Рисунок 3.11 Вікно з попередженням про неповноту даних

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.

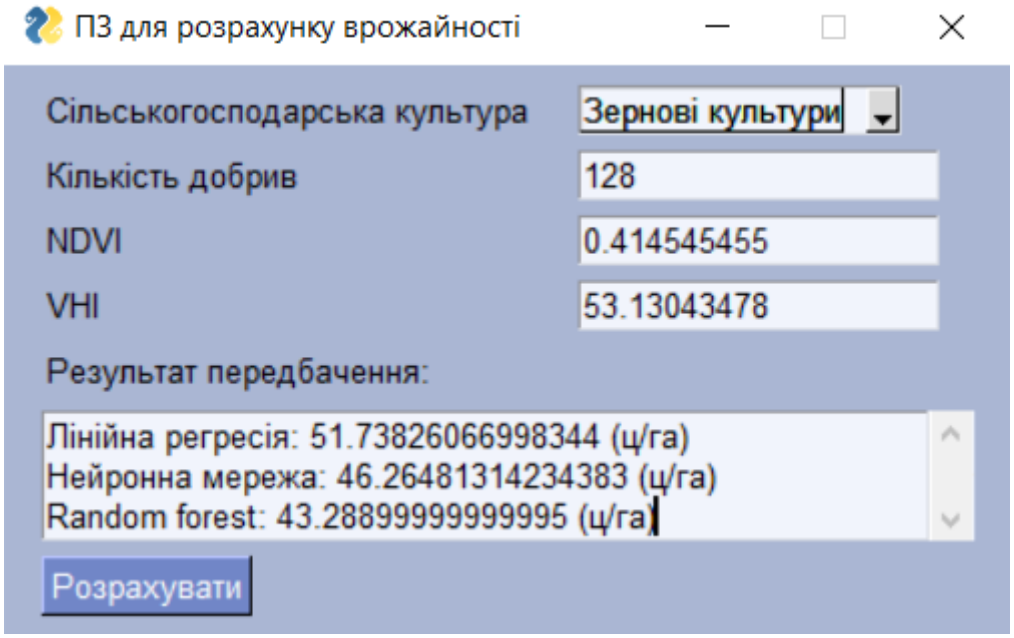

Рисунок 3.12 Результати прогнозування для зернових культур

Модель лінійної регресії передбачила врожайність на 2018 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, ~~в~~ 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 цукровий буряк. Реальний показник врожайності у 2018 році склав 509 ц/га.

Результати прогнозування врожайності цукрового буряку наведені на рис. 3.13.

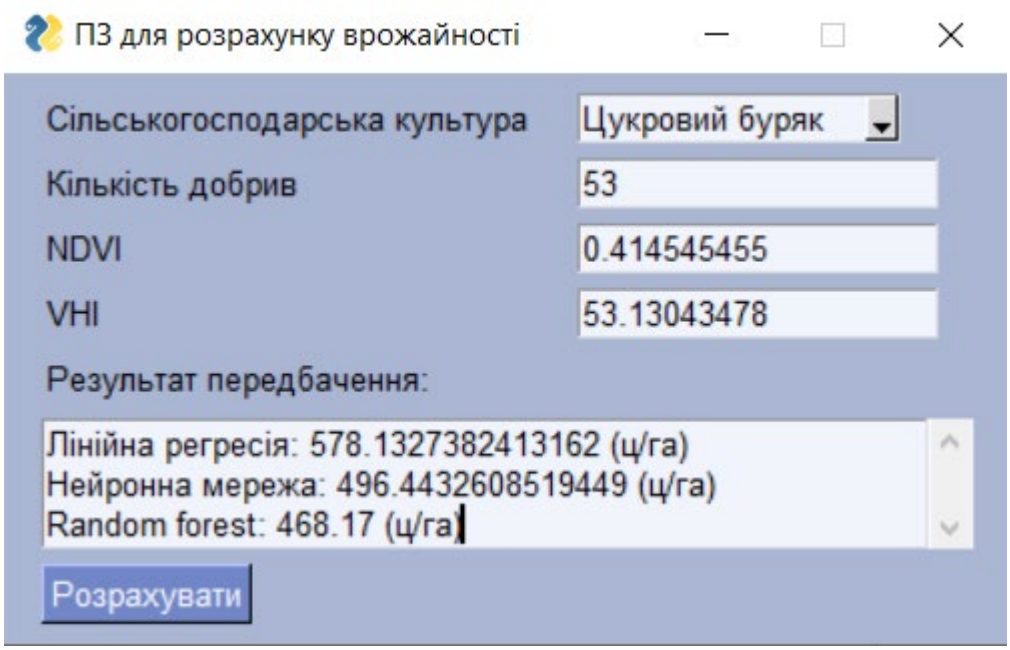

Рисунок 3.13 Результати прогнозування для цукрового буряку

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 соняшник. Реальний показник врожайності у 2018 році склав 23 ц/га.

Результати прогнозування врожайності соняшнику наведені на рис. 3.14.

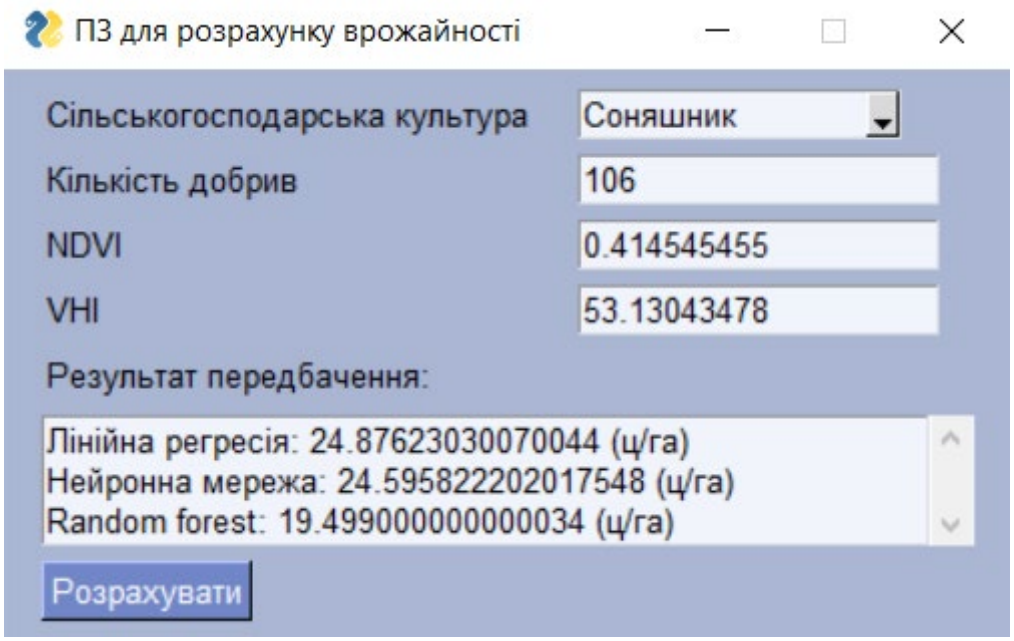

Рисунок 3.14 Результати прогнозування для соняшнику

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	Цукровий буряк	Соняшник
	Реальне значення (ц/га)	47,4	509	23
Лінійна регресія	Передбачене значення (ц/га)	51,74	578	24,9
	Похибка (%)	9,2	13,6	8,2
Нейронна мережа	Передбачене значення (ц/га)	46,26	496	24,6
	Похибка (%)	2,4	2,6	7

Random forest	Передбачене значення (ц/га)	43,29	468	19,5
	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 методу Random forest дали найбільше значення mse порівняно з моделями на інших методах, на зернових культурах та цукровому буряку, але одне з найменших на соняшнику. Проте на етапі експерименту вони дали протилежний результат: прогнозування зернових культур та цукрового буряку дало меншу похибку, порівняно з моделями лінійної регресії, але модель для прогнозування врожайності соняшнику дала найбільшу похибку.

ВИСНОВКИ

У магістерській роботі здійснено теоретичне узагальнення та вирішення науково-прикладної проблеми щодо використання методів машинного навчання в прогнозуванні врожайності декількох основних сільськогосподарських культур на території України.

На основі отриманих теоретичних та практичних результатів зроблено такі висновки:

1. Проведено аналітичний огляд літератури щодо сучасного стану питання використання методів машинного навчання в прогнозуванні врожайності, як вітчизняних, так й іноземних науковців. На основі цього огляду було обґрунтовано актуальність обраної теми та виконано постановку задачі.
2. Описано декілька методів машинного навчання, які можна використати для розв’язання поставлених задач, а саме поліноміальну регресію, нейронні мережі та Random forest. Для кожного методу було наведено математичні основи та описано основні параметри й алгоритми, які вони використовують.
3. Розглянуто та описано різні види вхідних даних, які можна застосувати для прогнозування врожайності сільськогосподарських культур. Особливу увагу було приділено показникам NDVI та VHI. Обґрунтовано доцільність вибору таких вхідних даних для розв’язання поставленої задачі.
4. Створено вибірку з описаними вхідними даними для кожної сільськогосподарської культури, які вибрані для дослідження. Проведено попередній аналіз отриманих даних методом візуалізації та кореляційного аналізу. На основі створеної вибірки виконано оцінку точності прогнозування декількох варіантів архітектури, описаних раніше, методів машинного навчання. Для отримання оцінки використано метод перехресної перевірки з MSE. На основі отриманої оцінки, обґрунтовано вибір архітектури для побудови кінцевого програмного забезпечення.

5. Розроблено програмний додаток, який використовує моделі на основі машинного навчання для прогнозування врожайності сільськогосподарських культур на території України. Для створення програмного додатку, проведення аналізу даних та перехресної перевірки використано мову високого рівня Python та набір бібліотек: Numpy, Seaborn, Pandas, Scikit-learn, Pickle, PySimpleGUI.
6. Проведено експериментальне прогнозування врожайності сільськогосподарських культур на 2018 рік. У результаті експерименту було отримано похибку прогнозу для кожного методу машинного навчання та кожної сільськогосподарської культури. Середня похибка для лінійної регресії склала 10,3%, нейронної мережі – 4%, Random forest – 10,6%.
7. На основі аналізу отриманих експериментальних результатів показано, що для обраних сільськогосподарських культур та вхідних даних найоптимальнішими являються моделі на основі нейронних мереж, які дали набагато точніший результат порівняно з іншими методами. При виборі інших вхідних даних можна отримати інший результат, що підтверджується попередніми науковими публікаціями.
8. За результатами експерименту було виявлено розбіжності між оцінкою точності прогнозування перехресної перевірки та експерименту, що представляє 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.html28. 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. GitHub. Machine-learning-octave – Режим доступу до електронного ресурсу: https://github.com/trekhleb/machine-learning-octave/tree/master/neural-network

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

32. 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.

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

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

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

36. 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.

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

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

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

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

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

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

43. NOAA STAR Center for satellite application and research:

https://www.star.nesdis.noaa.gov/

44. 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/

45. 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

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

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

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

49. 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.

50. 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

51. 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.

52. 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'] == 'Цукровий 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'] == 'Соняшник':
```

```
        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()
```