НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені Ігоря СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ФІЗИКО-ТЕХНІЧНИЙ
ІНСТИТУТ
Кафедра математичного моделювання та аналізу даних

«До захисту допущено»
В.о. завідувача кафедри
_____Іван ТЕРЕЩЕНКО
«__»_____2023 р.

Дипломна робота

на здобуття ступеня бакалавра

зі спеціальності: 113 Прикладна математика

на тему: **«Методи аналізу вартості земельних ділянок за супутниковими та геопросторовими даними»**

Виконав: студент 4 курсу, групи ФІ-92
Марінченко Микита Олександрович
Керівник: доктор філософії, с. в. кафедри ММАД _____
Яйлимова Г.О.
Рецензент: старший викладач Наконечна Ю.В. _____

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ — 2023

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені Ігоря СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ФІЗИКО-ТЕХНІЧНИЙ
ІНСТИТУТ
Кафедра математичного моделювання та аналізу даних

Рівень вищої освіти — перший (бакалаврський)

Спеціальність (освітня програма) — 113 Прикладна математика,

ОПП «Математичні методи моделювання, розпізнавання образів та комп'ютерного зору»

«До захисту допущено»
В.о. завідувача кафедри
_____Іван ТЕРЕЩЕНКО
«__»_____2023 р.

ЗАВДАННЯ
на дипломну роботу

Студент: Марінченко Микита Олександрович

1. Тема роботи: «Методи аналізу вартості земельних ділянок за супутниковими та геопросторовими даними»,

керівник: доктор філософії, с. в. кафедри ММАД Яйлимова Г.О.,

затверджені наказом по університету №__від «__»_____2023 р.

2. Термін подання студентом роботи: «__»_____2023 р.

3. Вихідні дані до роботи: *опубліковані джерела за тематикою дослідження.*

4. Зміст роботи: *досліджуючи класичні та сучасні методи оцінки земельних ділянок, за допомогою виключно об'єктивних супутникових даних наданих у відкритому доступі, використовуючи геопросторові характеристики земельних ділянок на території України, побудовано математичну модель апроксимації нормалізованої грошової оцінки*

вищезазаначених земельних ділянок, за основу тренування було взято оцінки цих земельних ділянок експертами за 2021-2022 роки.

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо): презентація доповіді.

6. Дата видачі завдання: 15 листопада 2022 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання	Примітка
1	Узгодження теми роботи із науковим керівником	15-30 листопада 2022 р.	Виконано
2	Огляд опублікованих джерел за тематикою дослідження	Листопад-Грудень 2022 р.	Виконано
3	Пошук та обробка бази даних	Січень-лютий 2023 р.	Виконано
4	Розробка моделі по актуальним даним	15 січня - 10 лютого 2023 р.	Виконано
5	Планування ітерацій моделі	Лютий 2023 р.	Виконано
6	Програмна реалізації моделі	Березень 2023 р.	Виконано
7	Експериментальні ітерації моделі	15 - 30 березня 2023 р.	Виконано
8	Корегування параметрів та оптимізація найкращої ітерації	10 - 20 квітня 2023 р.	Виконано
9	Обробка результатів та висновки	10 - 20 травня 2023 р.	Виконано
10	Оформлення дипломної роботи	Травень 2023 р.	Виконано

Студент _____ Марінченко М.О.

Керівник _____ Яйлимова Г.О.

РЕФЕРАТ

Обсяг роботи 65 сторінкок, містить 1 таблицю та 14 рисунків. Для дослідження було використано 12 бібліографічних найменувань.

Ціна земельних ділянок – це вічно актуальна частина економіки, яка завжди потребує великої уваги. Неважливо яка ваша ціль придбання земельної ділянки, завжди потрібно дивитися на деякі об’єктивні фактори які будуть впливати на її ринкову оцінку.

Багато хто, банально не знаючи як користуватися доступними ресурсами, вибирає шмат землі який “сподобався” і наймає експерта за немалу суму грошей для того щоб просто її оцінити та через деякий час дати своєму клієнту оцінку даної ділянки. Однак якщо знати які головні фактори впливають на оцінку землі, особливо ті які дуже легко знайти у відкритому доступі, та вивчити як їх використати, можна дуже швидко знайти приблизну оцінку будь-якої ділянки. Це зекономить і час, і гроші, що дуже важливо для оптимального використання свого капіталу на кожному вільному ринку.

Датасет, який використовується для роботи взятий з офіційного державного сайту України land.gov.ua у розділі “Моніторинг Земельних Відносин”. У ньому присутні дані зі супутникових знімків території України, і ціни деяких з цих земель на час 2021-2022 року. Буде розроблятись та використовуватись модель яка буде тренуватись знаходити залежності та закономірності у даних датасету які дозволять їй зробити приблизну об’єктивну оцінку земної ділянки.

Внесок цієї дипломної роботи полягає у створенні швидкого та стабільного методу пошуку оцінки земної ділянки за об’єктивними супутниковими даними у відкритому доступі або отриманими вручну. Що дозволить обробляти більшу кількість інформації перед експертною оцінкою ділянки, що займає час та потребує фінансів.

Ключові слова: машинне навчання, дані , грошова оцінка, тренування моделі, перетворення даних, валідація даних.

SUMMARY

The price of land plots is an ever-relevant part of the economy that always requires great attention. Regardless of your goal in acquiring a land plot, it is always necessary to consider certain objective factors that will influence its market valuation. Many people, simply unaware of how to use available resources, choose a piece of land that they 'like' and hire an expert for a significant amount of money just to evaluate it and eventually provide their client with an assessment of the plot. However, if you know the main factors that affect land valuation, especially those easily found in open access, and learn how to utilize them, you can quickly find an approximate valuation for any plot. This saves both time and money, which is crucial for optimal use of capital in any free market.

The dataset used for this work is taken from the official government website of Ukraine, land.gov.ua, in the section 'Monitoring of Land Relations.' It contains data from satellite images of the territory of Ukraine and the prices of some of these lands during the period of 2021-2022. A model will be developed and used, which will be trained to find dependencies and patterns in the dataset that will allow it to make an approximate objective assessment of a land plot.

The contribution of this thesis lies in the creation of a fast and stable method for searching for land plot evaluations based on objective satellite data available in open access or obtained manually. This will enable processing a larger amount of information before the expert assessment of the plot, which takes time and requires finances.

Keywords: machine learning, data, monetary valuation, model training, data transformation, data validation.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

НГО – Нормативна Грошова Оцінка

ЕГО – Експертна Грошова Оцінка

ВСТУП

Актуальність роботи:

У сучасному світі проблеми пов'язані з оцінкою вартості земельних ділянок є надзвичайно важливими для економіки, будівництва, землекористування та розвитку міст і сільських територій. Потреба в точних і об'єктивних методах оцінки земельних ділянок виникає з метою прийняття ефективних рішень щодо їх використання, купівлі-продажу, оренди та розвитку. Однак, оцінка вартості земельних ділянок є складним завданням, яке вимагає комплексного підходу та використання сучасних методів аналізу.

З огляду на постійне зростання попиту на земельні ресурси та їх неймовірну складність придбання та утилізації, існує потреба в розробці методів аналізу, які б забезпечували об'єктивні та точні оцінки їх вартості. В цьому контексті використання супутникових та геопросторових даних стає надзвичайно важливим інструментом, який дозволяє отримати об'єктивну інформацію про земельні ділянки і здійснити їх грошову оцінку.

Застосування супутникових та геопросторових даних у поєднанні з методами аналізу може забезпечити об'єктивну оцінку вартості земельних ділянок. Витрачений час на оцінювання та порівняння значно скоротиться, особливо для користувачів з обмеженим бюджетом.

Мета дослідження:

Метою даної дипломної роботи є дослідження методів аналізу вартості земельних ділянок за допомогою супутникових та геопросторових даних. Основною метою є розробка швидкого та стабільного методу оцінки, який забезпечує точність і ефективність процесу. Розробка моделі яка дозволить це робити швидко та ефективно.

Завдання дослідження:

- Аналіз наукової літератури щодо методів оцінки вартості земельних ділянок та використання супутникових та геопросторових даних.
- Вивчення та аналіз доступних супутникових та геопросторових даних, що стосуються земельних ділянок.
- Розробка методики обробки та аналізу супутникових та геопросторових даних для оцінки вартості земельних ділянок.
- Тренування та валідація моделі на основі зібраних даних для отримання об'єктивних оцінок вартості земельних ділянок.

Об'єкт дослідження:

Земельні ділянки, супутникові та геопросторові дані.

Предмет дослідження:

Просторові та економічні зв’язки, які будуть впливати на ефективність та актуальність оцінки.

Методи дослідження:

Використання математичного аналізу та статистики, обробка даних, методи навчання моделі.

Новизна роботи:

Полягає в розробці методу оцінки вартості земельних ділянок, який може бути використаний в реальних умовах для прийняття обґрунтованих рішень щодо землекористування, купівлі-продажу та розвитку міст і сільських територій виключно на основі супутникових даних. Результати дослідження можуть бути корисні для оціночних компаній, землевпорядних органів, розвитку міст та регіонального планування, та приватних осіб які потребують швидшу оцінку більшої кількості даних.

РОЗДІЛ 1.
АНАЛІЗ ВАРТОСТІ ЗЕМЕЛЬНИХ ДІЛЯНОК ТА МЕТОДІВ ОЦІНКИ

1.1 Оцінка земної ділянки

1.1.2 Методи оцінки

Оцінка ринкової ціни земельної ділянки включає в себе ряд методів, які використовуються для визначення поточної вартості земельного об'єкта.

Основні методи оцінки ринкової ціни земельної ділянки включають:

1) Порівняльний підхід: Цей метод базується на порівнянні земельної ділянки з аналогічними ділянками, що продавалися на ринку. Оцінювач аналізує ринкові ціни і характеристики подібних ділянок, такі як розташування, розмір, форма, доступність до комунікацій та інші фактори. За допомогою цього порівняння визначається вартість оцінюваної ділянки.

2) Вартість заміщення: Цей метод базується на вартості будівництва аналогічної земельної ділянки. Вартість будівництва включає затрати на придбання землі, розробку, планування, будівництво та інші витрати. Оцінювач враховує ці затрати і використовує їх як основу для визначення ринкової ціни ділянки.

3) Дохідний підхід: Цей метод застосовується, коли земельна ділянка використовується з метою генерації доходу, наприклад, в оренду або для комерційної діяльності. Оцінювач аналізує потенційний дохід, який може бути здобутий з використання земельної ділянки, і використовує методи, такі як капіталізація доходу або рентабельність, для визначення ринкової ціни.

4) Інвестиційний аналіз: Цей метод використовується для оцінки земельної ділянки як інвестиційного активу. Оцінювач аналізує фінансові показники, такі як внутрішній прибуток (IRR) або назад до окупності (ROI), для визначення потенційного прибутку, який може

бути отриманий від інвестиції в ділянку. Цей метод часто використовується для комерційних та розроблюваних земельних ділянок.

У будь-якому випадку необхідно включити інформацію про фізичні характеристики об'єкта і його правовий статус. Враховуючи регіональні особливості, де знаходиться земельна ділянка, а також її близькість до населеного пункту, оскільки ці фактори мають вплив на ринкову вартість.

Усі ці методи враховують наміри покупця щодо майбутнього придбання земельної ділянки, або її соціоекономічні властивості, які дуже сильно відрізняються в залежності від локального ринку. Однак для початку треба знайти які фактори дають нам саме найкращу ділянку по характеристикам, які потім можна буде прикладати до більш локальних ринків, оскільки вони є фундаментальними у оцінюванні будь-якої ділянки.

1.1.3 Види оцінки

В залежності від мети та методів проведення, оцінка земель поділяється на різні види:

1) Бонітування ґрунтів: Цей вид оцінки зосереджений на визначенні якості ґрунту та його придатності для сільськогосподарського виробництва. Бонітування оцінює потенціал ґрунту для вирощування різних сільськогосподарських культур та враховується при визначенні екологічної придатності земель.

2) Економічна оцінка земель: Цей вид оцінки використовується для визначення економічної цінності земельних ділянок. Він включає аналіз використання земель порівняно з іншими ресурсами, ефективності господарювання на земельних ділянках та економічної придатності земель для різних цілей.

3) Грошова оцінка земельних ділянок: Цей вид оцінки визначає ринкову вартість земельних ділянок у грошовому виразі. Існують два підвиди грошової оцінки: НГО та ЕГО.

Хоча бонітування ґрунтів та економічна оцінка земель є важливими методами оцінки землі, у нашому випадку їх використання може бути менш доцільним для тренування моделі, яка базується на супутникових та геопросторових даних.

Бонітування ґрунтів зазвичай базується на фізичних властивостях ґрунту, таких як структура, текстура, вологість і т.д. Економічна оцінка земель, зі свого боку, залежить від факторів, пов'язаних зі споживачами земельних ділянок та ринковими умовами. Ці методи не мають прямого зв'язку з геопросторовими характеристиками, які можна легко отримати з супутникових даних.

Бонітування та економічна оцінка земель часто базуються на експертній оцінці та суб'єктивних критеріях. Це може включати оцінку фахівцями, експертні групи або економічні моделі. У такому випадку, навчання моделі на основі цих методів буде складним, оскільки потрібно мати чіткі та об'єктивні вхідні дані для тренування.

Вони можуть вимагати додаткових досліджень, аналізу даних, збору інформації про фізичні, екологічні, соціальні та економічні аспекти. Це є часо-й трудомістким процесом, який ускладнить розробку та тренування моделі.

1.1.4 Грошова оцінка земельних ділянок

1.1.4.1 Нормативна грошова оцінка

НГО земельної ділянки виконується за допомогою професійних оцінювачів або експертів, які мають спеціалізовані знання і досвід в галузі оцінки нерухомості.

Першим кроком є збір усієї необхідної інформації про земельну ділянку, такої як розташування, розмір, призначення, фізичні характеристики,

правовий статус та інші важливі фактори. Ці дані можуть бути отримані з офіційних реєстрів, кадастрових агентств, документації власника та інших джерел.

Для проведення оцінки використовуються нормативні параметри, встановлені законодавством. Ці параметри можуть включати ставки податків, орендної плати, норми рентабельності, втрат виробництва та інші фактори, які впливають на вартість земельної ділянки.

На основі зібраних даних і нормативних параметрів проводиться розрахунок ринкової вартості земельної ділянки. Цей процес може включати використання статистичних методів, порівняльного аналізу з аналогічними ділянками, врахування економічного потенціалу та інші підходи для визначення ціни.

Залежно від конкретної мети оцінки і правового статусу земельної ділянки, проводиться встановлення остаточної вартості. Це може включати додаткові корекції або врахування факторів, таких як особливості ринку, регіональні чи місцеві фактори, відсоткові ставки та інші обставини, що можуть впливати на вартість.

1.1.4.2 Експертна грошова оцінка

Ций тип оцінки земельної ділянки виконується за допомогою професіонального експерта або оцінювача, який має необхідну кваліфікацію і досвід в галузі оцінки нерухомості.

Процес експертної грошової оцінки включає збір усієї необхідної інформації про земельну ділянку, включаючи фізичні характеристики, розташування, розмір, правовий статус, доступність інфраструктури та інші фактори, які можуть впливати на вартість ділянки.

Експерти вибирають відповідну методику оцінки в залежності від мети оцінки та характеристик ділянки. Методики можуть включати порівняльний аналіз з аналогічними ділянками, дохідний підхід, вартість заміщення або

комбінацію різних підходів. Далі проводять аналіз і розрахунки, використовуючи обрані методи оцінки. Це може включати порівняння з аналогічними ділянками, розрахунок доходності, врахування вартості будівель, інфраструктури, розрахунок земельних резервів та інші показники, що впливають на вартість.

На основі проведених розрахунків і аналізу експерти визначають остаточну вартість земельної ділянки. Це може бути вартість на основі ринкової вартості, реконструйованої вартості, зведеної вартості тощо. Вартість земельної ділянки визначається з урахуванням всіх факторів, що впливають на неї.

1.1.5 Вибір грошової оцінки

У цій роботі дотепніше вибрати НГО земельної ділянки. Нормативна грошова оцінка базується на встановлених законодавством нормах, правилах та методиках. Це означає, що оцінка виконується згідно зі стандартами, які є загальними для всіх оцінювачів і експертів. Такий підхід забезпечує більш об'єктивні результати, оскільки враховуються загальноприйняті критерії та методи.

Вона має чітко визначені методи та алгоритми, які використовуються для визначення вартості земельних ділянок. Це забезпечує прозорість процесу оцінки та уніфікацію підходів. Ми використаємо ці методи для тренування своєї моделі, оскільки вони мають чіткі кроки та правила, які можна відобразити у вигляді алгоритмів.

Найголовніший фактор для цієї роботи це факт того що, на відміну від експертної, нормативна грошова оцінка зазвичай базується на загальнодоступних даних, таких як земельні кадастрові дані, правовий статус ділянки, зонування, нормативи тощо. Це означає, що можна легко отримати необхідні дані для тренування моделі та перевірки її результатів.

Обравши нормативну грошову оцінку та її методи для тренування моделі, ми маємо можливість оцінити вартість земельних ділянок з урахуванням загальноприйнятих стандартів та правил. Це дозволить отримати об'єктивні та перевірені результати, які будуть корисні для подальшого дослідження та практичного застосування у оцінці ділянок.

1.1.6 Процес нормативної оцінки

Починається зі збору вихідних даних про земельну ділянку, яку планується оцінювати. Це може включати кадастровий номер, площу ділянки, її розташування (координати), призначення землі, наявність будівель або споруд, правовий статус, земельні обмеження та інші важливі характеристики. Цю інформацію можна отримати з кадастрових баз даних, органів місцевого самоврядування, реєстрів нерухомості та інших відповідних джерел.

Далі, збирається ринкова інформація про подібні земельні ділянки у відповідному регіоні. Це може включати дані про ціни на схожі ділянки, орендні ставки, цінові тенденції, дані про транзакції продажу землі та інші ринкові фактори, які впливають на вартість землі. Для цього можна скористатися базами даних, статистичними звітами, оголошеннями про продаж землі та іншими джерелами.

Головна частина цього процесу для нас це оцінка фізичних характеристик земельної ділянки. Фактори, як рельєф, ґрунти, дренаж, доступ до комунікацій, екологічні аспекти та інші характеристики, які можуть впливати на вартість землі. Для збору цих даних можуть використовуватися геодезичні дослідження, геологічні звіти, аерофотозйомка, супутникові знімки та інші методи.

Перевіряється правовий статус земельної ділянки та виявляється наявність будь-яких обмежень, обтяжень або правових спорів. Це може включати аналіз документації про право власності, межових документів, існуючих обмежень, дозвільних процедур та іншого відповідного юридичного матеріалу.

Досліджується наявність та якість інфраструктури та сервісів, таких як дороги, електромережі, водопостачання, каналізація, газопостачання та інші, може також вплинути на вартість земельної ділянки. Аналіз даних про наявність та якість інфраструктури та сервісів, які доступні на ділянці або у безпосередній близькості до неї.

Після збору всіх вихідних даних вони потребують систематизації та обробки. Це включає упорядкування даних, видалення аномалій, розрахунок показників, створення бази даних і підготовку для подальшого аналізу.

Наступним кроком є застосування нормативних методів оцінки, які визначаються відповідними нормативними актами або регламентами. Ці методи використовуються для визначення вартості земельної ділянки на основі зібраних і проаналізованих даних. Вони можуть включати використання нормативних коефіцієнтів, таблиць, алгоритмів або моделей, які визначаються у відповідних документах.

Застосовуючи нормативні методи, проводиться оцінка вартості земельної ділянки.

1.1.7 Актуальність супутникових даних

Супутникові дані надають різноманітні характеристики земної ділянки, які можуть бути використані для аналізу вартості та оцінки. Їх можна розділити на декілька видів:

1) Географічні:

Відомості про ділянку, такі як її положення, розмір, форму, контур і орієнтацію. Ці дані можуть бути використані для визначення меж ділянки, її конфігурації та інших геометричних параметрів.

2) Ландшафтні:

Рельєф та особливості ділянки, такі як нахил, наявність водних джерел, лісові масиви, рівень розташування відносно моря тощо. Ці характеристики можуть впливати на вартість земельної ділянки, наприклад, в разі наявності панорамного виду або природних резерватів.

3) Використання землі:

Поточне використання земельної ділянки, таке як види діяльності (сільське господарство, промисловість, житлове будівництво тощо) та ступінь їх інтенсивності. Це дозволяє оцінити економічну придатність ділянки та визначити потенціал для майбутнього розвитку.

4) Покриття землі:

Луки, сільськогосподарські поля, ліси, міська забудова, водні об'єкти тощо. Це важливо при визначенні призначення земельної ділянки та врахуванні екологічних факторів.

5) Зміни в часі:

Історична інформація про зміни, що відбулися на ділянці протягом певного періоду. Це дає змогу аналізувати динаміку змін у використанні землі та розвитку навколишнього середовища.

З тих даних які підходять для використання у цій моделі є саме географічні дані. Незалежно від намірів використання, саме фізичні характеристики дають безупереджену оцінку, яка є першопочатковою базою на основі якої будуються усі інші плани щодо використання ділянки, і власне додаткові фактори ринкової ціни.

Деякі геопросторові дані можуть також грати роль соціоекономічних – наприклад, якщо ділянка знаходиться в надзвичайно зручному місці, близько до міста або важливих інфраструктурних об'єктів, то це значно підвищує її цінність. В той же час, віддаленість від основних доріг чи недоступність до комунікаційної інфраструктури сильно знижують вартість ділянки.

1.2 Датасет та робота з даними

1.2.1 Пошук та вибір датасету

Вибір датасету це фундаментальне рішення для всієї подальшої роботи. Супутникові дані зазвичай є тільки на офіційних або визнаних джерелах через вартість використовуваного обладнання.

Так звані приватні особи та менш локальні компанії можуть тягнути за собою деякі ускладнення у використанні їх даних. Тому краще шукати у державних джерелах України, це дає деякі переваги.

Наприклад, державні ресурси зазвичай забезпечують доступ до високоякісних і достовірних даних. Використання офіційних джерел дозволяє уникнути потенційних проблем, пов'язаних з якістю, точністю та достовірністю даних.

Ці ресурси зазвичай оновлюють свої дані регулярно. У нашому випадку датасет оновлюється кожен рік. Поточний стан земельних ділянок дуже важливий у більш точній оцінці ділянки.

Також ці дані підпадають певним правовим нормам і мають чіткі умови використання, що дозволяє спокійно використовувати їх у дослідженні.

Репрезентативнівність є фактором який також не потрібно ігнорувати, бо велика кількість даних про різні області дає можливість тренувати модель давати більш об’єктивну та актуальну оцінку, незалежно від суто локальних факторів.

Мій вибір – це датасет взятий з land.gov.ua у розділі “Моніторинг Земельних Відносин”, у якому буде проходити подальша робота.

1.2.2 Робота з датасетом

1.2.2.1 Перетворення та валідація даних

Після того як датасет вибраний необхідно завантажити датасет у програмне середовище, у нашому випадку Python.

Перед перетворенням даних знайомимось зі структурою та змістом датасету. Дослідження стовпців, їхніх типів, значень та взаємозв'язків допоможе зрозуміти, які перетворення можуть бути необхідними.

Перетворення та валідація даних для вилучення некоректних даних з датасету є важливим етапом при нашому аналізі. Цей процес включає в себе ряд дій, спрямованих на перевірку, корекцію та вилучення некоректних,

неповних або недостовірних даних з датасету, щоб забезпечити точність та достовірність аналізу.

Одним з перших кроків є перетворення даних у відповідний формат. Це може включати конвертацію різних типів даних, наприклад, перетворення дати та часу в стандартний формат, переведення координат у потрібну систему координат, а також нормалізацію даних для забезпечення однорідності та порівнянності.

Після перетворення даних проводиться валідація, що полягає у перевірці правильності та достовірності даних. Вона має декілька етапів:

1. Перевірка на відповідність формату: перевірка, чи відповідають дані встановленому формату (наприклад, чи є числовими дані, які повинні бути числами).
2. Перевірка на відсутність дублікатів: перевірка, чи містить датасет дублікати записів. Дублікати можуть виникати внаслідок помилок при зборі даних або при обробці.
3. Виявлення відхилень: пошук аномальних значень або викидів, які можуть вказувати на помилки в даних або незвичайні явища. Наприклад, це може бути надмірно висока або низька вартість, що не відповідає загальній вибірці.
4. Зведення до єдиного джерела: перевірка, чи всі дані у датасеті мають посилання на відповідні джерела або документи. Це дозволяє перевірити автентичність та достовірність даних.
5. Заповнення пропущених значень: якщо у датасеті є пропущені дані, можна використати різні методи для їх заповнення. Наприклад, це може бути середнє значення, усе залежно від контексту та характеру даних.
6. Валідація логічних залежностей: перевірка, чи відповідають дані логічним залежностям або обмеженням. Наприклад, вартість будь-якої земельної ділянки не може бути від'ємною або нульовою.

При перетворення та валідації даних, некоректні та недостовірні дані можуть бути вилучені або позначені як сумнівні для подальшого аналізу. Це допомагає забезпечити точність та достовірність результатів аналізу і дозволяє уникнути неправильних висновків на основі неправдивої інформації.

1.2.2.2 Класифікація даних

Класифікація даних це групування даних у певні категорії або класи відповідно до їх властивостей, характеристик або особливостей. Класифікація даних дозволяє визначити закономірності, зробити прогнози та зробити висновки на основі аналізу.

Визначення класів це перший крок у класифікації даних. Він вимагає визначення категорій або класів, на які будуть поділені дані. В аналізі вартості земельних ділянок це можуть бути класи вартості, типи землі і тд.

Для класифікації даних необхідно вибрати ознаки або змінні, за якими будуть проводитись розподіл та класифікація. В аналізі вартості земельних ділянок це можуть бути географічні координати, площа ділянки тощо. Перед застосуванням методів класифікації необхідно підготувати дані. Це може включати видалення аномалій, заповнення пропущених значень, стандартизацію даних або вибір підмножини ознак, які найбільше впливають на класифікацію.

Ну і оцінка - після застосування методів класифікації необхідно оцінити їх ефективність і якість. Це може включати порівняння прогнозних результатів з реальними мітками, використання метрик оцінки якості (наприклад, точність, чутливість, специфічність).

Після проведення класифікації даних можна використовувати результати для аналізу вартості земельних ділянок. Наприклад, можна визначити вартісні класи земельних ділянок на основі їх характеристик, встановити взаємозв'язки між вартістю та іншими змінними.

Класифікація даних є потужним інструментом. Вона дозволяє розподілити дані у відповідні категорії, виявити закономірності, зробити

прогнози та зробити обґрунтовані висновки на основі аналізу класифікованих даних.

1.2.2.3 Нормалізація даних

Процес нормалізації даних є одним з методів масштабування числових ознак у датасеті. Нормалізація дозволяє привести значення ознак до специфічного діапазону або шкали, забезпечуючи стандартизацію і стабільність даних для подальшого аналізу та моделювання.

Найпоширенішим методом нормалізації є Min-Max scaling, який перетворює значення ознак таким чином, щоб вони потрапляли в певний діапазон, наприклад, від 0 до 1. Процес нормалізації за методом Min-Max scaling може бути описаний наступним чином:

1) Визначення діапазону: Спочатку необхідно визначити діапазон, до якого ми хочемо перетворити значення ознак. Зазвичай це діапазон від 0 до 1, але його можна змінити в залежності від вимог та характеристик даних.

2) Обчислення мінімума та максимума: Знаходяться мінімальні та максимальні значення ознак у датасеті. Це можна зробити для кожної ознаки окремо або взяти загальні мінімальне та максимальне значення для всього датасету.

3) Формула нормалізації: Для кожного значення ознаки застосовується формула нормалізації, яка перетворює його до діапазону [0, 1]. Формула має наступний вигляд:

$$normalized\_value = (value - min\_value) / (max\_value - min\_value)$$

де:

normalized_value - нормалізоване значення ознаки;

value - початкове значення ознаки;

min_value - мінімальне значення ознаки;

max_value - максимальне значення ознаки.

4) Застосування нормалізації: Для кожного значення ознаки застосовується формула нормалізації з кроку 3, щоб отримати нормалізовані значення для всього датасету.

Перетворені значення ознак після нормалізації будуть знаходитись в заданому діапазоні і матимуть стандартизований розподіл, що полегшує порівняння та аналіз даних.

Нормалізація даних особливо корисна, бо наші алгоритми вимагають стандартизованих даних.

1.2.3 Опис даних

Опис даних є важливою частиною аналізу даних і досліджень, яка має на меті надати інформацію про структуру, властивості і характеристики набору даних.

Він надає основні статистичні показники, такі як середнє значення, медіана, мода, мінімум, максимум, стандартне відхилення, квартилі та інші характеристики, які допомагають зрозуміти центральну тенденцію, розкид і форму розподілу даних.

Опис даних може містити інформацію про якість та достовірність даних. Це може включати перевірку на пропущені значення, викинуті або викидні значення, аномальні спостереження та інші неперевірені аспекти даних. Якісний опис даних допомагає зрозуміти, наскільки надійні та повні дані і як це може вплинути на подальший аналіз і результати.

Для кожної змінної або ознаки можуть бути надані детальні характеристики, такі як тип даних, одиниці виміру, можливі значення, інтерпретація та розуміння значень. Це допомагає зрозуміти сутність і значення кожної змінної та спрощує інтерпретацію результатів аналізу. Це є важливим етапом підготовки та розуміння даних перед їх подальшим аналізом.

1.3 Вибір та конфігурація моделі

1.3.1 Вибір моделі згідно до задачі

Першим кроком є вибір моделей, які підходять для задачі. Це можуть бути методи машинного навчання, статистичні моделі, нейронні мережі або геостатистичні методи.

1.3.1.1 Класифікація задачі

Передбачання вартості земельних ділянок є пошуком залежності між вхідними змінними (супутникові та геопросторові дані) і вихідною змінною (вартість земельних ділянок). У цій задачі, вартість земельних ділянок є неперервним числовим значенням, що потрібно передбачити на основі інших змінних.

Таким чином, можна використати регресійні методи для побудови моделі, яка апроксимує залежність між супутниковими та геопросторовими даними і вартістю земельних ділянок. Це можуть бути методи, такі як лінійна регресія, дерева рішень, випадковий ліс або нейронні мережі, які можуть знаходити складні залежності між змінними та передбачати числові значення вартості земельних ділянок на основі вхідних даних. Отже, найкращий метод підходу до цієї задачі є метод регресії.

1.3.1.2 Переваги нейронної мережі

Нейронні мережі мають потужність моделювання, яка дозволяє виявляти та ураховувати складні нелінійні зв'язки між вхідними даними та вартістю земельних ділянок. Вони виявляють нелінійні закономірності, які можуть бути пропущені менш складними моделями, такими як лінійна регресія.

Задача аналізу вартості земельних ділянок включає велику кількість ознак, таких як параметри супутникових зображень, геопросторові дані, економічні показники тощо. Нейронна мережа здатна ефективно

опрацьовувати інформацію з великої кількості ознак та автоматично виявляти релевантні залежності, що можуть покращити точність прогнозування вартості.

Нейронні мережі можуть бути гнучкими та адаптивними до змін у даних та завданнях. Вони можуть самостійно навчатися та підлаштовуватися до нових закономірностей у даних без необхідності вручну налаштовувати модель. Це особливо корисно в задачах аналізу земельних ділянок, де залежності можуть змінюватися з часом та варіювати в залежності від географічного розташування.

Застосування нейронних мереж дозволяє ефективно обробляти великі обсяги даних. Завдяки паралельним обчисленням та розподіленому навчанню, нейронні мережі можуть ефективно працювати з великими супутниковими та геопросторовими наборами даних, що є важливим у задачі аналізу вартості земельних ділянок.

Нейронні мережі можуть враховувати контекстуальну інформацію про земельні ділянки, таку як прилеглість до інших об'єктів, транспортну доступність, інфраструктуру тощо. Вони можуть аналізувати ці дані та використовувати їх для кращого прогнозування вартості.

1.3.1.3 Прототип моделі для вирішення регресивної задачі

Враховуючи регресивну природу задачі, можна розглянути різні моделі, такі як лінійна регресія, дерева рішень, випадковий ліс тощо.

Використовуючи навчальний набір даних, модель буде навчатися на основі алгоритмів регресії та оптимізації. Метою навчання є знаходження оптимальних ваг та параметрів моделі для прогнозування вартості земельних ділянок.

Після навчання моделі використовується тестовий набір для оцінки її ефективності та точності. За допомогою метрик оцінки, таких як середня квадратична помилка (Mean Squared Error), коефіцієнт детермінації (R-

squared) та інші, можна визначити, наскільки добре модель прогнозує вартість земельних ділянок.

Після успішного навчання та оцінки моделі, її можна використовувати для прогнозування вартості земельних ділянок на нових невідомих даних.

1.3.2 Основні проблеми застосування нейронних мереж

Застосування нейронних мереж відкриває широкі можливості для розв'язання складних завдань у багатьох галузях, але воно також пов'язане з деякими проблемами.

Нейронні мережі зазвичай потребують великої кількості даних для ефективного навчання. Обробка великого обсягу даних може вимагати значних обчислювальних ресурсів, включаючи сильний процесор та велику кількість оперативної пам'яті. Це може становити проблему, особливо якщо вам необхідно працювати з обмеженими обчислювальними ресурсами.

Вони мають багато параметрів, які потребують налагодження та оптимізації під час навчання. Велика кількість параметрів може призводити до перенавчання моделі, коли вона добре пристосовується до тренувальних даних, але показує погану здатність узагальнювати на нові дані. Це вимагає обережного контролю та налагодження параметрів моделі.

Вимога значної кількості даних для ефективного навчання. У випадку, коли у вас обмежений обсяг доступних даних, може бути складно використовувати нейронні мережі. Недостатньо даних може призводити до недофіксування моделі та недостатньої здатності до узагальнення.

Нейронні мережі часто вважаються "чорним ящиком", оскільки їх внутрішній процес вирішення проблеми не завжди легко пояснити або інтерпретувати. Вони можуть давати точні прогнози, але можуть бути важкими в поясненні причин, що стоять за цими прогнозами. Це може бути проблемою, особливо якщо потрібно пояснити прийняті рішення або зробити висновки на основі моделі.

Навчання складних нейронних мереж може вимагати значної обчислювальної потужності та тривалого часу. Використання глибоких нейронних мереж з багатьма шарами та нейронами може займати багато годин або навіть днів для завершення навчання. Це може створювати перешкоди при експериментах та швидкому розгортанні моделей.

Застосування нейронних мереж може стикатися з деякими проблемами, але їх можна запобігти або пом'якшити їх вплив.

Недостатня кількість даних може призвести до перенавчання моделі або низької загальної відтворюючої здатності. Щоб запобігти цьому, необхідно збирати достатньо об'єктів для навчання моделі. Якщо дані обмежені, треба використовувати методи регуляризації, такі як зменшення кількості параметрів моделі або використання розширених методів навчання на незначній кількості даних.

Перенавчання відбувається, коли модель стає надто складною та специфічною для навчальних даних, і не може загальноюзагальнювати на нові дані. Цьому можна запобігти за допомогою регуляризації, наприклад, застосуванням методу Dropout або обмеженням глибини архітектури мережі. Також важливо використовувати достатньо валідаційних даних та перевіряти модель на них під час навчання.

Нейронні мережі потребують повних та безперервних наборів даних для ефективного навчання. Якщо деякі дані відсутні, можна використовувати методи заповнення пропусків даних, такі як середнє значення, медіана або інтерполяція. Також можна використовувати методи зневаження (ignoring) пропущених даних у випадках, коли вони є незначними для аналізу.

Вони можуть бути обчислювально вимогливими, особливо для складних моделей або великих наборів даних. Щоб зменшити витрати обчислювальних ресурсів, можна використовувати апроксимаційні методи, такі як пришвидшення навчання, оптимізація гіперпараметрів, використання менших архітектур мережі або використання розподіленого навчання на кластері ресурсів.

Нейронні мережі є складними у виведенні зрозумілих та інтерпретованих результатів. Щоб вирішити цю проблему, можна використовувати методи важливості ознак, які дозволяють зрозуміти, які фактори внесли найбільший вклад у прогнози моделі. Також можна використовувати більш прості моделі для порівняння з нейронними мережами та забезпечення більшої інтерпретованості результатів.

1.3.3 Регресія у машинному навчанні

Неважливо наскільки відпрацьована та натренована модель, оцінка точності передбачень цієї моделі та ефективність алгоритму це головний орієнтир результату, для цього є дві важливі метрики, які потрібно враховувати: дисперсія та зміщеність.

Дисперсія вказує на різницю в оцінці цільової функції, якщо б використовувалися різні навчальні дані. Ця цільова функція встановлює залежність між вхідними та вихідними змінними. При використанні інших наборів даних, цільова функція повинна залишатись стабільною з мінімальними змінами, оскільки модель повинна бути однаковою для будь-якого типу даних.

Щоб уникнути неправильних передбачень, необхідно забезпечити низьку дисперсію. З цією метою модель повинна бути генералізованою, щоб приймати невидимі характеристики і коригувати передбачення.

Зміщеність - це стійка тенденція алгоритму навчатися неправильним залежностям, не враховуючи всю доступну інформацію в даних. Для отримання точних результатів модель повинна мати низьку зміщеність. Наявність невідповідностей у наборі даних, таких як відсутні значення, недостатня кількість даних або помилки вхідних даних, призводить до високої зміщеності і неправильних передбачень.

Точність і помилка - це дві важливі метрики. Помилка вимірює розбіжність між фактичним значенням і прогнозованим значенням, яке модель

оцінює. Точність вказує на відношення правильних передбачень, зроблених моделлю.

Щоб модель була ідеальною, вона має мати низьку дисперсію, низьку зміщеність і низьку помилку. Для досягнення цього, набір даних розбивається на тренувальний і тестовий набори. Модель навчається на тренувальних даних, а її ефективність оцінюється на тестових даних. Для зменшення помилки під час навчання моделі застосовується функція помилки. Якщо модель просто запам'ятовує тренувальні дані, не розпізнаючи патерни, вона дає невірні прогнози для нових даних. Графік, побудований на основі такої моделі, буде проходити через всі точки даних, і точність на тестовому наборі буде низькою. Це явище називається перенавчанням і виникає через високу дисперсію.

З іншого боку, якщо модель працює добре на тестових даних, але з низькою точністю на тренувальних даних, це призводить до недонавчання.

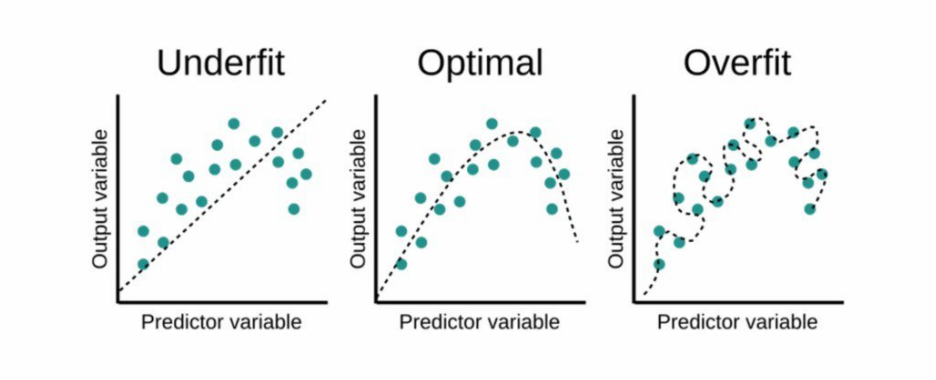

Рис. 1.3.3 Приклади недонавчання (underfit) та перенавчання (overfit) [7]

Існує низка різних алгоритмів для створення регресійної моделі, які будуть розглянуті далі.

1.3.4 Лінійна регресія

Лінійна регресія в машинному навчанні використовує математичні методи для знаходження лінійного зв'язку між залежною змінною та однією або декількома незалежними змінними. Це досягається шляхом знаходження найкращої прямої лінії, яка найкраще підходить до навчальних даних.

У лінійній регресії, для здійснення прогнозу, модель обчислює зважену суму значень вхідних ознак, до якої додається константа, відома як зсув (bias). Залежна змінна в цьому випадку є неперервною, а незалежні змінні можуть бути як неперервними, так і дискретними. Лінія регресії представляє собою пряму лінію, яка найкраще апроксимує дані.

Математично, прогноз у лінійній регресії виражається рівнянням:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

Тут y - прогнозоване значення,

n - загальна кількість вхідних ознак,

x_n - значення n-ої вхідної ознаки,

θ_n - параметри моделі (θ_0 - зсув, а θ_1, θ_2, … θ_n - коефіцієнти).

Коефіцієнти визначають вагу, яку кожна вхідна ознака має у формуванні прогнозу. Зсув відповідає за положення лінії на вісі y і дозволяє краще підлаштовуватись під дані.

У лінійній регресії намагаються знайти такі коефіцієнти, які найкраще відображають залежність між змінними, щоб зробити точні прогнози.

Зміщення і дисперсія завжди знаходяться взаємозалежно. Високий рівень зміщення спостерігається, коли модель є дуже простою з обмеженою кількістю параметрів, тоді як низький рівень зміщення відповідає складній моделі з великою кількістю параметрів. Ми прагнемо до досягнення як найменшого зміщення, так і найменшої дисперсії. Для цього потрібно розбиратися з залежністю уважно, щоб отримати бажаний результат.

1.3.5 Проста лінійна регресія

Проста лінійна регресія - це простий, але потужний метод регресії, який дозволяє передбачити значення вихідної змінної на основі однієї вхідної змінної. Це досягається шляхом побудови прямої лінії, яка найкращим чином

відповідає залежності між цими змінними. Наприклад, за допомогою простої лінійної регресії можна передбачити оцінку студента на основі кількості годин, які він витрачає на навчання.

Математично це представляється рівнянням:

$$y = mx + c$$

де x - незалежна змінна (вхід),

y - залежна змінна (вихід),

m - нахил (коефіцієнт),

c - перехоплення (інтерсепт).

Проста лінійна регресія дозволяє знайти оптимальну пряму лінію, яка найкращим чином відповідає залежності між вхідними та вихідними змінними. Це допомагає нам зрозуміти, як зміна вхідної змінної впливає на вихідну змінну.

Уявіть, що дані точки на графіку, і за допомогою простої лінійної регресії ми можемо побудувати пряму лінію, яка найкращим чином апроксимує ці дані.

1.3.6 Множинна лінійна регресія

Цей метод схожий на просту лінійну регресію, але має більше однієї незалежної змінної. Кожне значення незалежної змінної x пов'язане зі значенням залежної змінної y. Оскільки ми маємо багатовимірне представлення, найкраща лінія підгонки - це площина.

Математично це можна виразити таким чином:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Для прикладу, можна уявити, що ми хочемо передбачити, чи студент складе іспит. Враховується декілька вхідних змінних, таких як кількість годин, проведених на навчання, загальна кількість предметів і кількість годин сну

попередньої ночі. Оскільки є декілька вхідних змінних, використовується множинна лінійна регресія.

1.3.7 Мультиваріативна лінійна регресія

Мультиваріативна лінійна регресія, використовується для роботи з кількома вихідними змінними. Наприклад, якщо лікар хоче оцінити стан здоров'я пацієнта на основі зібраних зразків крові, діагноз включатиме прогнозування кількох значень, таких як артеріальний тиск, рівень цукру та рівень холестерину.

1.3.8 Поліноміальна регресія

Хоча лінійна регресійна модель може знаходити залежності в заданому наборі даних шляхом побудови простого лінійного рівняння, вона може бути недостатньо точною для складних даних. В таких випадках потрібно використовувати криві, які краще відповідають даним, ніж прямі лінії. Один з підходів - використання поліноміальної моделі. В цій моделі ступінь рівняння більше одиниці. Математично поліноміальна модель виражається так:

$$Y_0 = b_0 + b_1 x_1 + ...b_n x_n$$

де Y_0 - прогнозоване значення для поліноміальної моделі з коефіцієнтами регресії b_1 до b_n для кожної степені, і b_0 - зміщення.

Якщо ступінь n = 1, то поліноміальне рівняння є лінійним рівнянням.

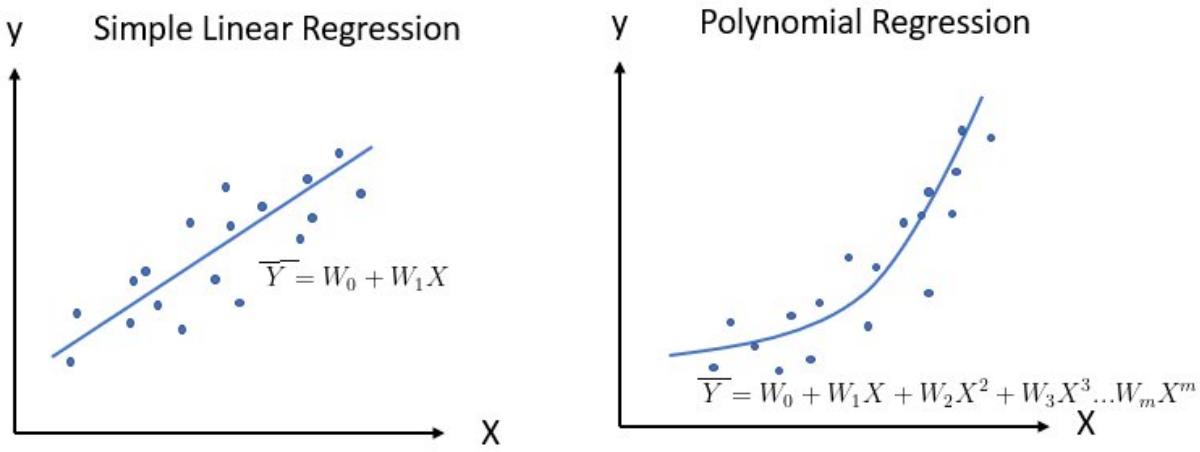

Рисунок 1.3.8. Проста лінійна регресія поряд з поліноміальною [8]

Щоб уникнути перенавчання, у машинному навчанні застосовуються методи регуляризації, такі як рідж-регресія і лассо-регресія, особливо коли враховується велика кількість ознак. Ці методи сприяють зменшенню ваги коефіцієнтів ознак і одночасно мінімізують різницю між прогнозованими і фактичними спостереженнями. Це допомагає уникнути перенавчання та забезпечити більш стабільні прогнози.

У рідж-регресії (регуляризації L2) до функції втрат додається штрафний член (λw_{i2}), де λ контролює рівень штрафу. Цей штраф допомагає зменшити значення коефіцієнтів ознак, запобігаючи їх занадто великому зростанню. Функція втрат:

$$J(w) = n_1(\sum i = 1_n(y\hat{}(i) - y(i))2 + \lambda w_{i2})$$

Коли $\lambda = 0$, ми повертаємося до перенавчання, а коли $\lambda = \infty$, вага стає занадто великою і виникає недонавчання. Тому параметр λ потрібно обирати ретельно, щоб уникнути обох цих ситуацій.

У лассо-регресії/регуляризації L1 додається абсолютне значення (λw_i) замість квадрату коефіцієнта. Це стоїть за "найменший вибірковий оператор звуження".

Тоді функція втрат буде:

$$J(w) = n_1(\sum i = 1_n(y^i - y(i))2 + \lambda w_i)$$

1.4 Функція активації

Це функція, яка використовується в штучних нейронних мережах і видає мале значення для невеликих вхідних значень і більше значення, якщо вхідні значення перевищують певний поріг. Якщо вхідні значення достатньо великі, функція активації "спрацьовує", в іншому випадку вона не реагує. Іншими словами, функція активації подібна до воріт, які перевіряють, чи є вхідне значення більше критичного числа.

Функції активації є корисними, оскільки вони додають нелінійність до нейронних мереж, дозволяючи їм вчитися потужні операції. Якщо функції активації були б видалені з прямого шарового нейронного мережі, всю мережу можна було б спростити до простої лінійної операції або матричного перетворення на вхідний сигнал, і вона б втратила здатність виконувати складні завдання, такі як розпізнавання зображень.

Відомі функції активації, що використовуються в науці про даних, включають функцію релу, сімейство сигмоїдних функцій, таких як логістична сигмоїдна функція, гіперболічний тангенс та арктангенс.

У галузі комп'ютерних наук функції активації були натхненні дією потенціалу дії в нейронах мозку. Коли електричний потенціал між внутрішньою і зовнішньою частинами нейрона перевищує певне значення, настає активація, і нейрон генерує сигнал, який передається сусіднім нейронам. Цей послідовний процес активації, відомий як "потік імпульсів", дозволяє сенсорним нейронам передавати відчуття від пальців до мозку і дозволяє руховим нейронам передавати команди від мозку до кінцівок.

1.4.1 Функція ReLU

У глибокому навчанні існує кілька широко використовуваних функцій активації. Одна з них - функція випрямленого лінійного вузла (ReLU), що має просту структуру. Вона повертає нуль, якщо вхідний сигнал є від'ємним, а в протилежному випадку без змін повертає сам вхідний сигнал.

$$f(x) = max(0,x)$$

Також варто розглянути обчислення градієнта функції ReLU, який математично не визначений при x = 0, але все ж корисний у нейронних мережах.

Нульовий градієнт для від'ємних значень x може створювати проблеми під час навчання нейронної мережі, оскільки нейрон може "застрягати" у нульовій області, і зворотне поширення не змінить його ваги.

Для $x > 0$: Похідна дорівнює 1.

Для $x \leq 0$: Похідна дорівнює 0.

У математичній нотації:

$$f(x) = 1, \text{якщо } x > 0$$

$$f(x) = 0, \text{якщо } x <= 0$$

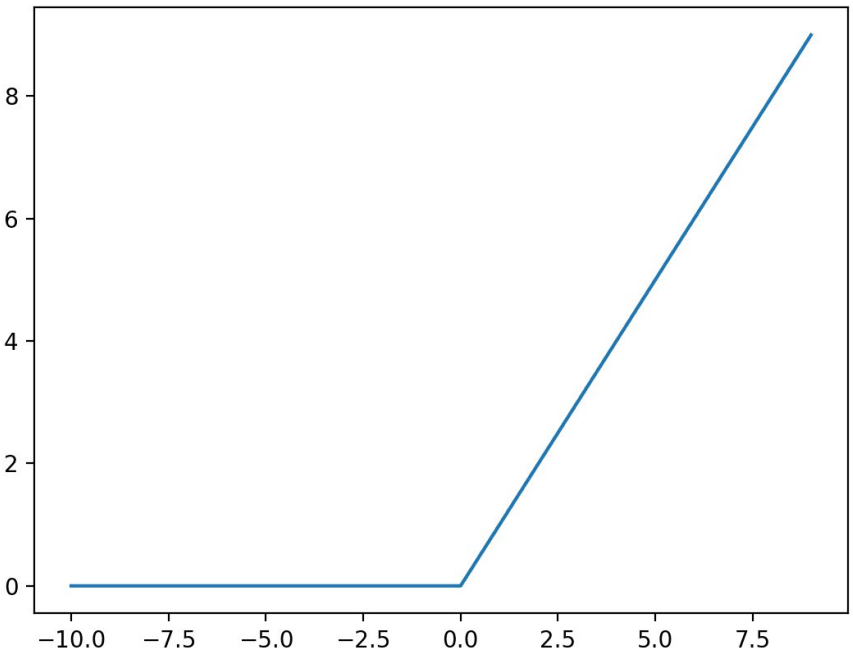

Рисунок 1.4.1. Функція ReLU [9]

1.4.2 Функція PReLU

У зв'язку з проблемою нульового градієнту, яку виникає у функції ReLU, часто застосовується модифікована версія, відома як параметричний прямокутний лінійний вузол (PReLU).

$$f(a,x) = \begin{cases} x, \text{якщо } x \geq 0 \\ ax, \text{якщо } x < 0 \end{cases}$$

Це має перевагу того, що градієнт відмінний від нуля в усіх точках (крім 0, де він не визначений).

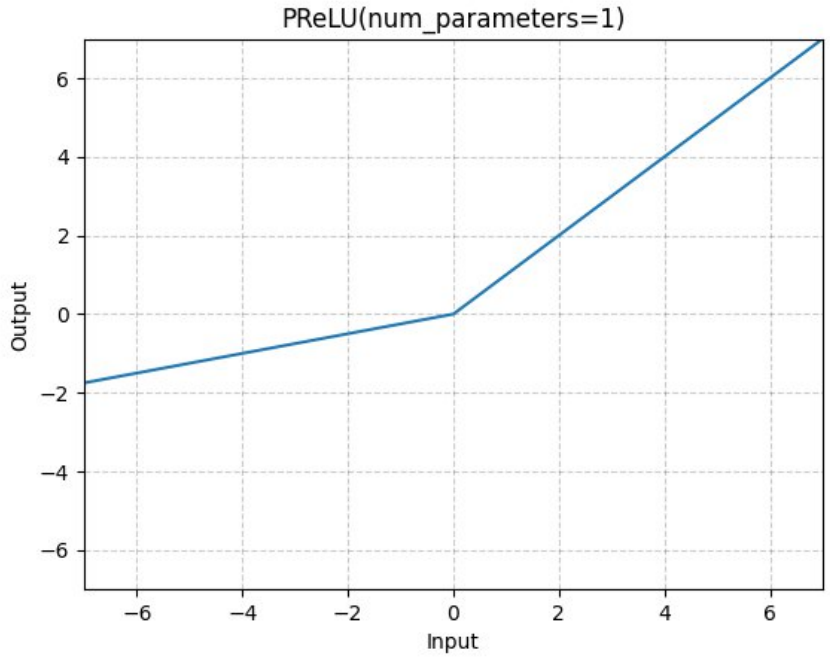

Рисунок 1.4.2 Функція PReLU [10]

1.4.3 Логістична Сигмоїдна Функція

Іншою відомою функцією активації є логістична сигмоїдна функція. Ця функція має те перевагу, що градієнт визначений у всіх точках (крім 0, де він не визначений), і її значення зручно знаходяться в діапазоні між 0 і 1 для будь-якого значення x. Математично визначена логістична сигмоїдна функція зручна для аналітичних розрахунків, але обчислення експоненційних функцій

може бути витратним з точки зору обчислювальної складності, тому в практиці часто використовуються більш прості функції, такі як ReLU:

$$f(x) = \frac{e^x}{e^x + 1}$$

Функція логістичної сигмоїди має корисну властивість - її градієнт визначений у всіх точках, і її значення зручно знаходяться в діапазоні від 0 до 1 для будь-якого значення x. Логістична сигмоїдна функція є математично зручною для обчислень, але обчислення експоненційних функцій можуть бути обчислювально витратними на практиці, тому простіші функції, наприклад ReLU, часто використовуються замість неї.

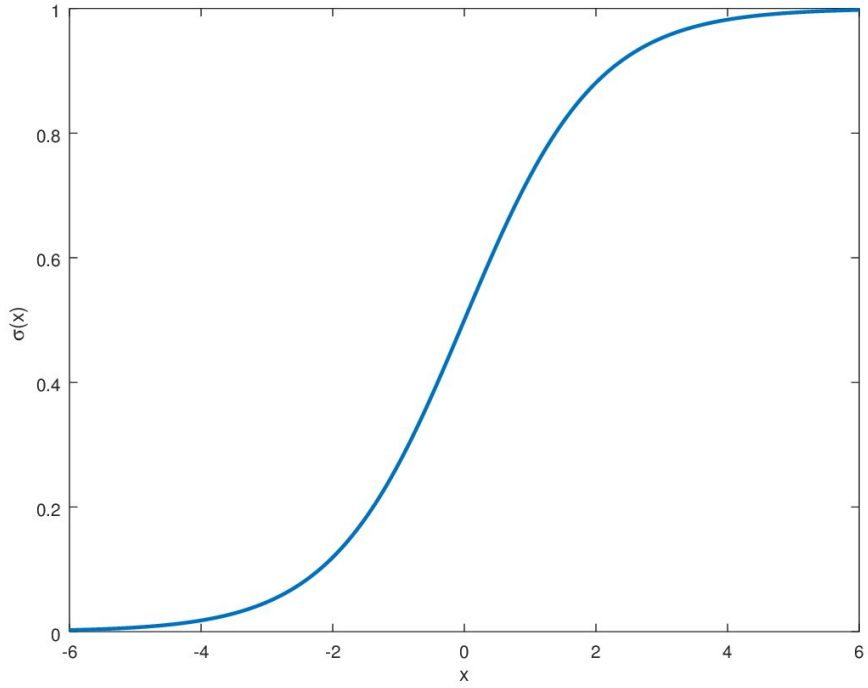

Рисунок 1.4.3. Логістична Сигмоїдна Функція [11]

1.4.4 Функція активації та Потенціал дії

Хоча ідея функції активації безпосередньо натхненна потенціалом дії в біологічній нейронній мережі, між цими двома механізмами є декілька відмінностей.

Біологічні потенціали дії не повертають неперервні значення. Біологічний нейрон або активується, або ні, схоже на функцію крокового порогу. Немає можливості для постійно змінної виходу, яка зазвичай необхідна для можливості зворотного поширення помилки.

Крім того, потенціал дії є функцією в часі. Коли напруга через нейрон перевищує порогове значення, нейрон спрацьовує і протягом наступних кількох мілісекунд передає характеристичну криву вихідної напруги своїм сусіднім нейронам, перш ніж повернутися до стану спокою.

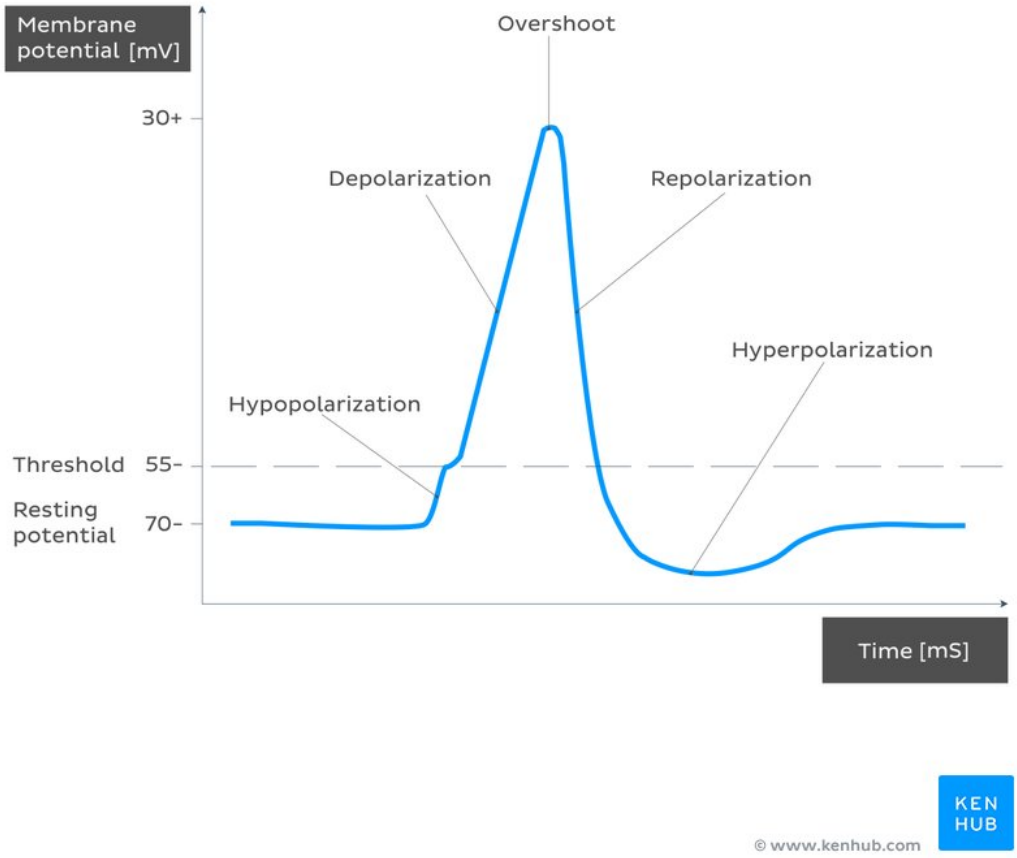

Рисунок 1.4.4. Потенціал дії нейрону відносно часу [12]

У людському мозку навчання відбувається асинхронно, кожен нейрон працює незалежно від інших та має зв'язок лише зі своїми найближчими сусідами. У штучних нейронних мережах "нейрони" зазвичай пов'язані з

тисячами або мільйонами інших нейронів та виконуються процесором в певному порядку.

Хоча біологічні нейрони не можуть повертати неперервні значення через часову компоненту, інформація передається за допомогою частоти спалахів та режиму спалаху нейрона.

Висновки до розділу 1

У першому розділі роботи були розглянуті теоретичні аспекти, необхідні для розуміння основних концепцій і методів аналізу вартості земельних ділянок з використанням супутникових та геопросторових даних.

Була проведена загальна характеристика оцінки земельної ділянки. Було розглянуто різні методи оцінки, включаючи грошову оцінку, яка зосереджується на визначенні фінансової вартості земельної ділянки. Зокрема, нормативна грошова оцінка та експертна грошова оцінка були розглянуті як основні підходи до визначення ціни земельних ділянок.

Описаний процес нормативної оцінки, що передбачає використання встановлених норм і правил для визначення грошової вартості земельної ділянки. Виявлено, що цей процес має важливе значення для урегулювання земельних відносин і забезпечення справедливої оцінки землі.

Показано актуальність використання супутникових даних у сфері аналізу вартості земельних ділянок. Супутникові дані надають значний обсяг інформації про земельну площу, місцевість та інші фактори, які можуть впливати на вартість землі. Це дозволяє більш точно визначати ціну земельних ділянок та проводити більш об'єктивні оцінки.

Розглянуті питання вибору та обробки даних. Вибір правильного датасету є важливим етапом у проведенні аналізу вартості земельних ділянок. Також було розглянуто методи перетворення, валідації, класифікації та нормалізації даних для підготовки їх до подальшого використання у моделі.

Розглянуто вибір та конфігурацію моделі для аналізу вартості земельних ділянок. Зокрема, показано, що нейронні мережі є потужним інструментом

для вирішення таких задач. Було розглянуто основні проблеми, пов'язані з застосуванням нейронних мереж, а також описано різні типи регресій, зокрема лінійну регресію і поліноміальну регресію.

Розглянуто функції активації, які використовуються у нейронних мережах для передачі сигналів між шарами.

Загальною метою розділу було ознайомлення з основними концепціями та методами аналізу вартості земельних ділянок з використанням супутникових та геопросторових даних. Висвітлені теоретичні відомості які надають необхідний фундамент для подальшого дослідження.

РОЗДІЛ 2.
ПРАКТИЧНЕ ЗАСТОСУВАННЯ МЕТОДІВ ОЦІНКИ У МОДЕЛІ

2.1 Датасет

2.1.1 Вибір датасету

З можливих варіантів вибору датасету було вирішено обрати датасет супутникових даних по геопросторовим характеристикам України, з офіційного вебсайту https://land.gov.ua, у розділі «Моніторинг земельних відносин».

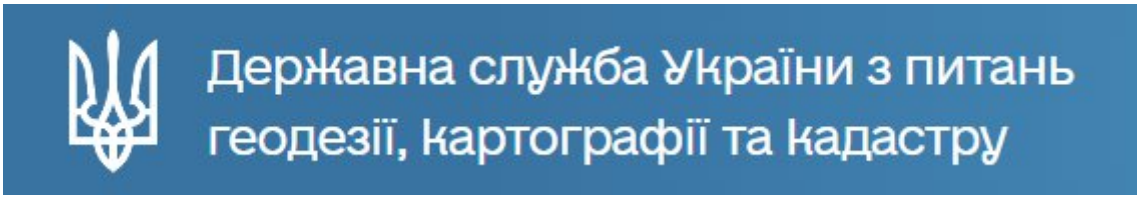

Рисунок 2.1.1. Скріншот сторінки

Таке джерело забезпечує доступ до вірогідних та авторитетних даних, пов'язаних зі земельними відносинами. Це має важливу роль у забезпеченні достовірності та надійності отриманих даних.

Державні сайти мають механізми оновлення та публікації актуальних даних. Використання актуального датасету дозволить отримати найсвіжішу інформацію про земельні ділянки, що є важливим для точного аналізу та оцінки їх вартості.

Це джерело надає широкий доступ до даних безкоштовно.Використання датасету з офіційного державного джерела дозволяє мати підтвердження та посилатися на відповідні правові норми та стандарти при аналізі та оцінці вартості земельних ділянок. Він містить інформацію про різні аспекти земельних ділянок, їх характеристики, стан власності, ринкову вартість та інші фактори. Такий обсяг даних надає можливість провести детальний аналіз і врахувати багатофакторні аспекти при визначенні вартості земельних ділянок.

2.1.2 Вибір даних

Вибір конкретних параметрів для аналізу вартості земельних ділянок є критичним кроком у дослідженні. Було обрано наступні параметри з датасету:

Площа поля (area): Площа земельної ділянки має прямий вплив на її вартість. Більша площа може вказувати на більший потенціал для сільськогосподарського виробництва або іншого використання, що позитивно впливає на вартість.

Нормалізований диференційний вегетаційний індекс за останній рік оцінки (ndvi_2021): Цей показник вказує на рівень розвитку рослинності на досліджуваній земельній ділянці. Він є індикатором якості грунту і потенціалу для сільськогосподарського виробництва.

Відстань до Києва (kyiv_dist): Розташування земельної ділянки відносно столиці сильно впливає на її вартість. Київ є важливим економічним центом з великими можливостями розвитку.

Відстань до найближчого міста (dist_citie): Врахування відстані до найближчого міста допоможе оцінити доступність і зручність для різних видів діяльності, а також вплинути на вартість земельної ділянки.

Відстань до найближчого населеного пункту (або містечка) (dist_town): Подібно до відстані до міста, врахування відстані до найближчого населеного пункту дає уявлення про доступність і потенціал розвитку землі.

Відстань до найближчого села (dist_villa): Розташування земельної ділянки відносно сіл може вплинути на її вартість, особливо якщо це сільськогосподарські землі або землі для інших сільськогосподарських потреб.

Відстань до найближчого елеватора (dist_eleva): Елеватори є важливими установками для зберігання і переробки сільськогосподарських продуктів. Близькість до елеватора впливає на доступність і зручність для сільськогосподарських діяльностей.

Відстань до найближчої головної дороги (dist_pr_pro), вторинної дороги (dist_sec_r) та сільської дороги (dist_count): Розташування земельної ділянки відносно доріг сильно впливає на доступність та транспортні зручності. Близькість до доріг робить ділянку більш привабливою для різних видів діяльності та вплинути на вартість.

Бонітет (bonitetl): Бонітет землі є показником її якості та плодючості. Високий бонітет вказує на високий потенціал для різних сільськогосподарських культур, що впливає на вартість.

Урожай основної культури (mjr_cr): Інформація про урожайність основної культури на земельній ділянці служить важливим фактором для визначення вартості. Високі урожаї свідчать про високий потенціал доходів, що позитивно впливає на вартість землі.

Тип покриття земної поверхні (LC_2021): Інформація про тип культур та загальне покриття земної поверхні на ділянці, такі як болото, ліс, поле тощо.

Передбачена ціна за гектар у 2021-2022 році (price): Цей параметр є цільовою змінною, передбачає модель.

Вибір декількох параметрів забезпечує баланс між точністю моделі, обчислювальною складністю і зрозумілістю результатів, що робить його ефективним підходом для роботи.

area	NDVI_2021	Kyiv_dist	Dist_Citie	Dist_Towns	Dist_Villa	Dist_Eleva
4,312399864	5,389631185	440,036	11,738	18,956	0,166	16,069
4,883900166	4,631513751	356,826	19,796	22,368	0,453	6,083
2,818300009	5,175996835	333,531	7,96	17,437	0,874	4,481
12,5	5,687799931	407,324	19,485	23,648	1,719	21,054
7,760399818	5,517038088	285,143	12,417	0,774	4,805	16,069

Таблиця 2.1.2

Приклад вибраних даних

2.1.3 Поділ датасету

Поділ даних на тренувальну, валідаційну і тестувальну вибірки є важливим кроком при розробці моделі машинного навчання. Цей підхід дозволяє оцінити ефективність моделі на незалежних даних і визначити, наскільки добре модель загальнізує знання з тренувальних даних на нових прикладах.

Тренувальна вибірка: Вона використовується для навчання моделі. Зазвичай це найбільша частина даних. Модель використовує ці дані для вивчення залежностей і підгонки параметрів.

Валідаційна вибірка: Ця вибірка використовується для налаштування гіперпараметрів моделі і оцінки її ефективності. Вона допомагає уникнути перенавчання моделі на тренувальних даних.

Тестувальна вибірка: Ці дані використовуються для оцінки остаточної ефективності моделі після завершення навчання і налаштування. Це незалежна вибірка, яка не використовувалась під час тренування та валідації.

2.1.4 Валідація та нормалізація даних

Усі дані у вибірці є числовими, і було реалізовано декілька можливих варіантів їх використання:

1) Якщо вхідні дані (наприклад, Kyiv_dist) містять помилку або неможливо перетворити їх на числове значення, то відповідний рядок не використовується у роботі. Це дозволяє уникнути використання неповних або некоректних даних, що можуть негативно вплинути на якість моделі.

2) Якщо вихідні дані містять помилку, їх виключають з усіх поділів датасету, але зберігають для майбутнього використання при передбаченні вартості. Такий підхід дозволяє уникнути впливу помилкових вихідних даних на тренування і оцінку моделі, але залишити можливість використання їх для подальших аналізів.

3) Якщо немає помилок ні в вхідних, ні в вихідних даних, то рядок використовується моделлю.

Такий підхід до використання даних дозволяє забезпечити якісну і надійну обробку інформації в моделі, уникнути неправильних або неповних даних, а також зберегти потенційно корисні дані для подальшого використання.

Коефіцієнти, такі як NDVI та MJR для нормалізації були поділені на максимальне значення, для того щоб перетворити їх у діапазон від одного до нуля, бо чим більше значення цих коефіцієнтів, тим дорожче ділянка.

Інші значення у вибірці це відстані, які були перетворені формулою, оскільки ціна від них обернено залежна:

$$n = \frac{a}{x + a}$$

де n - коефіцієнт;

x - відстань;

a – змінна, що характеризує швидкість падіння коефіцієнта.

Оскільки при нульовій відстані коефіцієнт дорівнює найбільшому можливому значенню (1), і при збільшенні відстані прямує до нуля.

2.1.5 Кореляція даних

Для роботи використовується мова Python, за допомогою pandas ми знаходимо кореляцію вибраних даних.

Кореляція дозволяє виявити наявність залежності між двома або більше змінними. Вартість земельних ділянок може бути вплинута різними факторами, такими як розмір ділянки, відстань до міста, якість землі тощо. Знаходження кореляційних зв'язків допомагає встановити, які змінні мають значущий вплив на вартість землі і як ці змінні взаємодіють між собою.

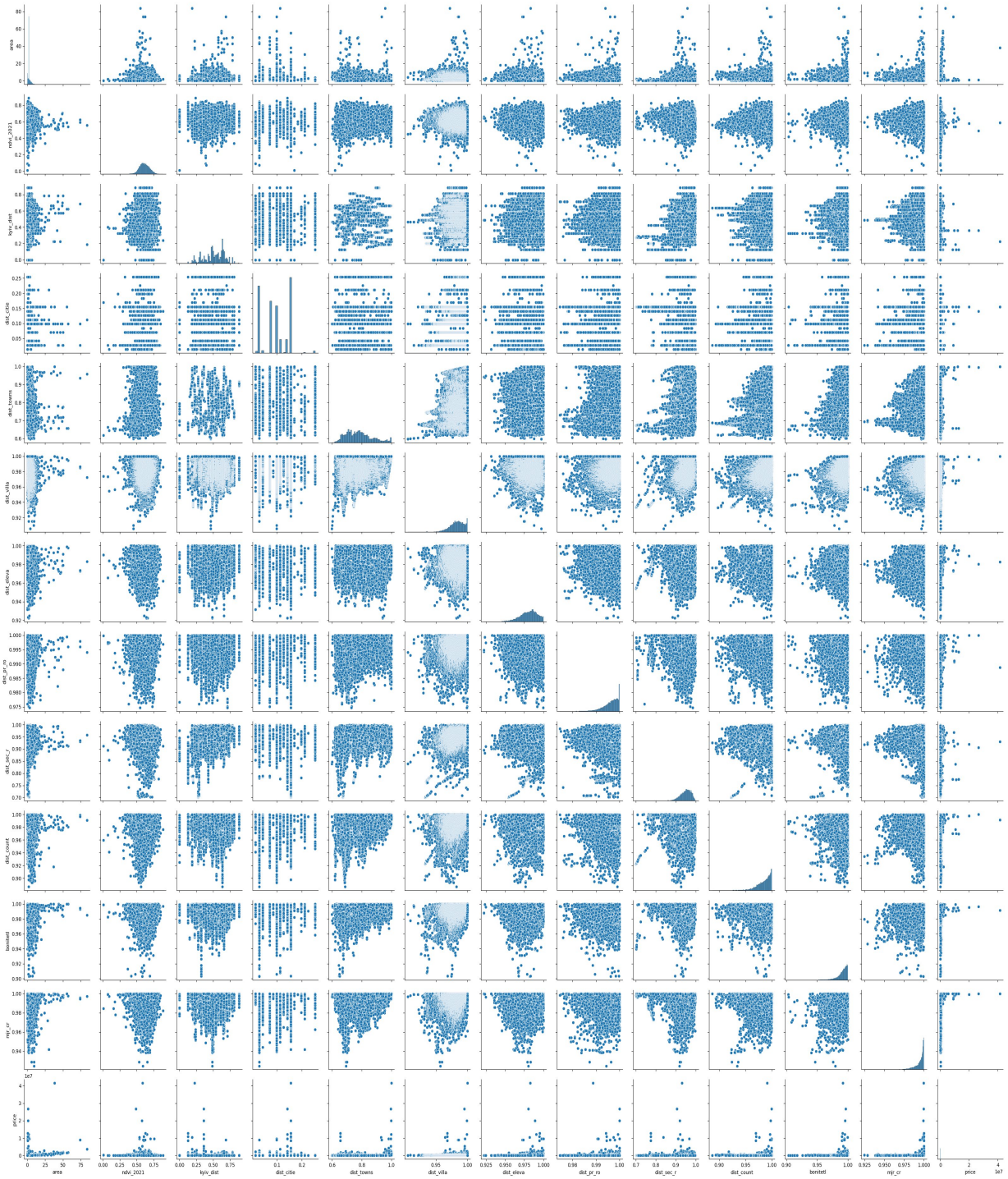

Рисунок 2.1.5.(1) Граф кореляції pairplot

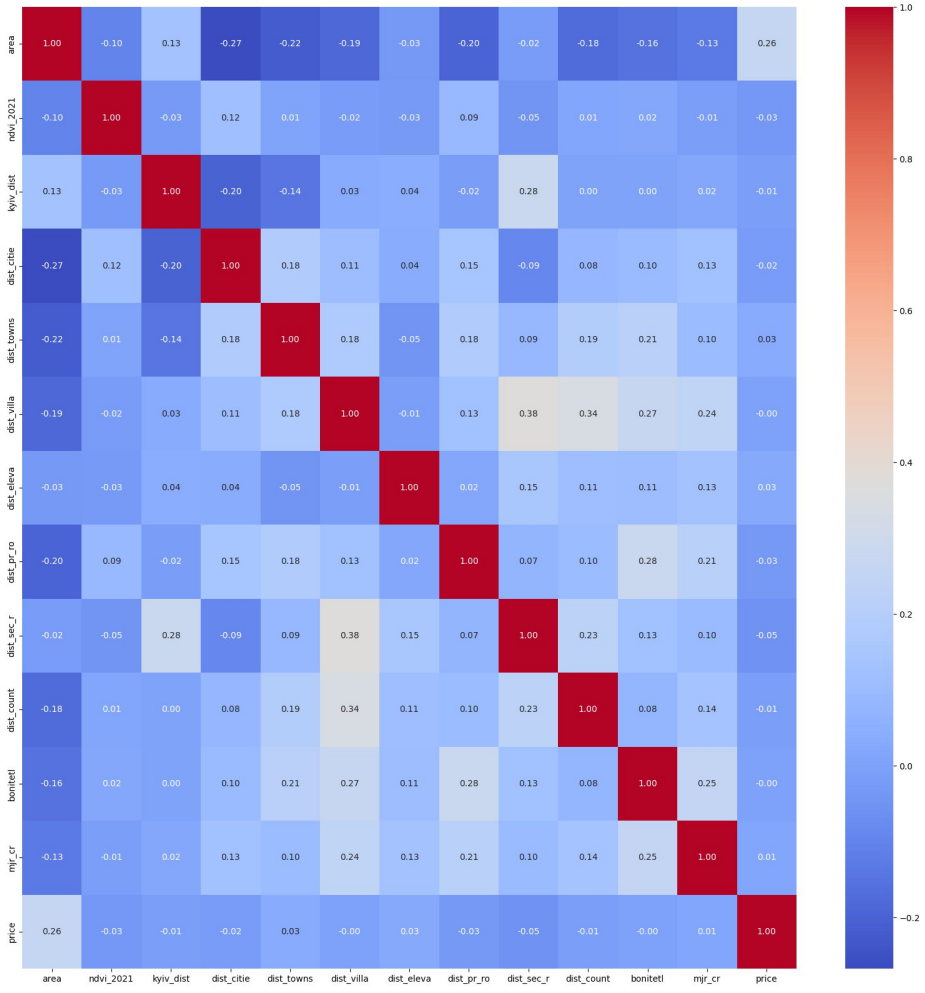

Рисунок 2.1.5.(2) граф кореляції heatmap

На жаль, рівень кореляції даних малий, тому це призводить до деяких проблем. Вона суттєво обмежує ефективність моделі з кількох причин.

Кореляція між змінними вказує на наявність залежностей та закономірностей між ними. Коли кореляція низька, це означає, що взаємозв'язок між змінними є слабким або його немає взагалі. В такому випадку модель може мати складнощі у передбаченні значень цільової змінної на основі незалежних змінних, оскільки вона не може виявити сильні впливи або взаємозв'язки між ними.

Низька кореляція між змінними означає, що вони майже незалежні одна від одної. У таких випадках модель може масти труднощі з встановленням чітких залежностей та здатністю передбачати значення цільової змінної. Це може призводити до підвищеної похибки прогнозування і менш точних результатів.

У випадку низької кореляції вибіркових даних, окремі змінні можуть мати обмежений вплив на цільову змінну. Це може позначитися на здатності моделі здійснювати точні прогнози або встановлювати значущі залежності між змінними.

Враховуючи ці фактори, низька кореляція даних у вибірці суттєво обмежує ефективність моделі, і може бути важко отримати точні та надійні результати.

2.2 Перша ітерація моделі

Перша ітерація моделі використовує тільки один внутрішній шар та функцію активатор SoftPlus.

1. Створюємо порожню модель за допомогою фреймворку машинного навчання TensorFlow.
2. Додаємо вхідний шар до моделі, вказуючи кількість вхідних змінних, які будуть використовуватись для тренування моделі. У нашому випадку це будуть обрані параметри, які були перераховані вище, наприклад, "area", "ndvi_2021", "kyiv_dist" тощо.
3. Додаємо внутрішній шар до моделі. У даному випадку це єдиний шар, і ми вибрали функцію активації SoftPlus для цього шару. Функція активації SoftPlus є нелінійною функцією, яка може допомогти моделі вивчити складніші залежності між вхідними та вихідними даними.
4. Додаємо вихідний шар до моделі. Вихідний шар має кількість нейронів, яка відповідає кількості параметрів, які ми хочемо передбачати. У даному випадку, може бути один нейрон.

Перед тренуванням моделі, необхідно виконати компіляцію, вказавши функцію втрати (наприклад, середньоквадратичну помилку для регресійної задачі), оптимізатор для навчання моделі та метрики, які будуть використовуватись для оцінки продуктивності моделі під час тренування.

За допомогою підготовленого датасету тренувальних даних, починаємо тренування моделі. Процес тренування включає передачу даних через модель, обчислення втрати, зворотнє поширення помилки та оновлення ваг моделі за допомогою градієнтного спуску. Цей процес повторюється на кілька епох для збільшення точності моделі.

Після завершення тренування моделі, ми можемо оцінити її продуктивність на валідаційному наборі даних. Це дозволяє нам визначити, наскільки добре модель виконує передбачення ціни за гектар.

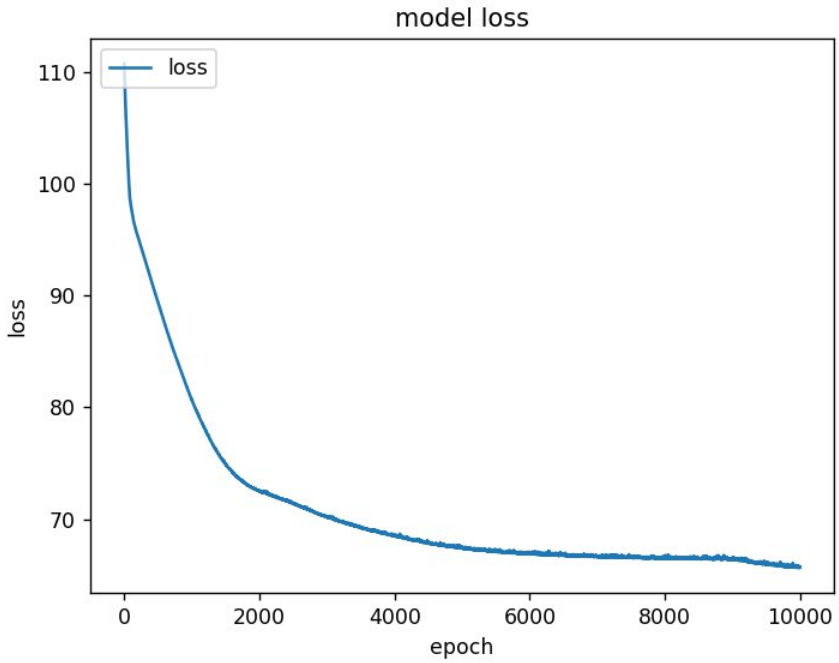

Рисунок 2.2 Граф ефективності моделі 1

Показник loss дає нам абсолютну похибку у відсотках відповідно до реальної ціни ділянки.

Очікувано, дивлячись на низьку кореляцію даних, модель дає передбачення з точністю усього 35%.

З огляду на низьку точність передбачення ітерації моделі, варто розглянути перехід до наступної моделі, яка може мати більшу складність та здатність адаптуватись до складних залежностей у даних.

2.3 Друга ітерація моделі

Друга ітерація моделі використовує два внутрішніх шари та функцію активатор ReLU. Вона є нелінійною функцією, яка дозволяє моделі набувати нелинійність та гнучкість у вирішенні завдань. Це дозволяє моделі більш точно моделювати складні залежності та розрізняти важливі функції у даних.

Після 10000 епох модель передбачає ціну вже краще, з точністю у 52%.

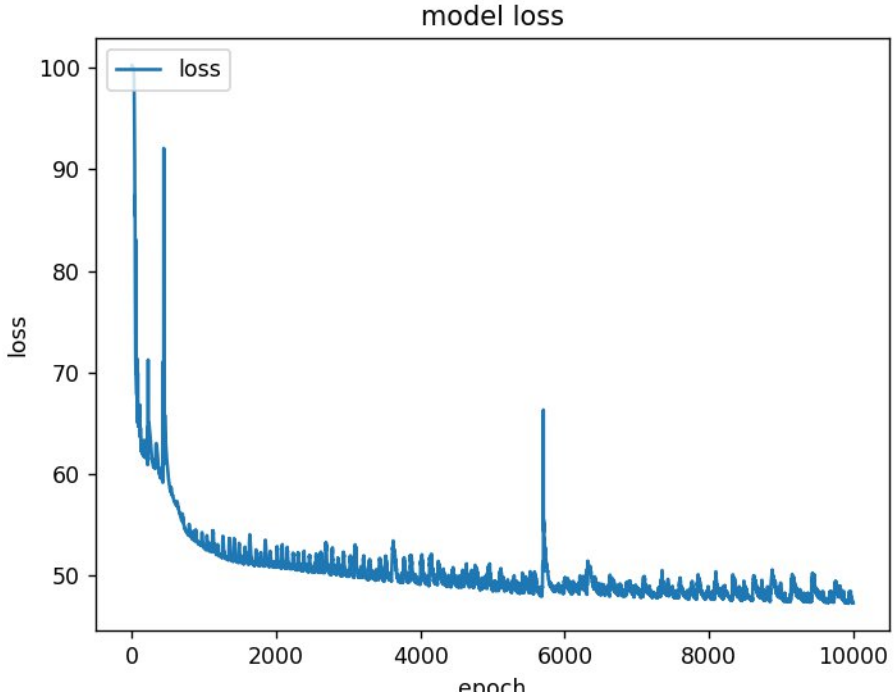

Рисунок 2.3. Граф ефективності моделі 2

Результати хоч і краще, однак все ще незадовільні і мають більш хаотичний характер. Додавання другого внутрішнього шару у модель призводить до збільшення складності моделі та кількості параметрів, які потрібно навчити. Це робить модель більш схильною до перенавчання. Це може пояснити хаотичні та нестабільні результати, де модель може

перестаратись підлаштуватись під окремі точки даних замість узагальнення. Однак зазвичай це вирішується довшим тренуванням.

2.4. Третя ітерація моделі

У третій ітерації моделі проведена спроба використати Random Forest.

Це ансамбль моделей рішень, що базуються на деревах прийняття рішень, який може використовуватися як для класифікації, так і для регресії.

З навчального набору даних випадковим чином вибираються підвибірки з поверненням (bootstrap samples). Це означає, що кожна підвибірка може містити деякі дані більше одного разу, а деякі дані можуть бути пропущені. Вибір таких підвибірок допомагає забезпечити різноманітність моделей рішень.

На кожній підвибірці будується окреме дерево прийняття рішень.

При побудові кожного дерева, тільки підмножина ознак використовується для поділу дерева. Це забезпечує більшу різноманітність і незалежність між деревами, оскільки різні ознаки використовуються для прийняття рішень.

Після побудови всіх дерев, для класифікації нового прикладу або передбачення значення регресії використовується голосування більшості або середнє значення прогнозів дерев.

```
forest_score: 0.3113103014594374
```

Рисунок 2.4. Результат роботи алгоритму Random Forest

На даний момент це найгірша точність з усіх ітерацій, у середньому 32%.

2.5 Четверта ітерація і основне пояснення ефективності

Остання модель також використовує активатор ReLU.

Перед початком роботи було розділено дані для тренування та тестування, у співвідношенні 2% тестових даних від усіх. Перед кожним

тестуванням дані для тренування також діляться на дані для тренування та дані для валідації, у співвідношенні 10%. Це потрібно, тому що під час тренування частина даних використовувалась лише для валідації для уникнення перенавчання. Після кінця тренування у програмі реалізована можливість виклику функції валідації на основі даних с першого поділу, які не використовувались ні для валідації, ні для тренування під час процесу тренування.

У ході роботи змінювалась конфігурація шарів моделі, найменшого значення втрат отримали при початковій конфігурації другої ітерації.

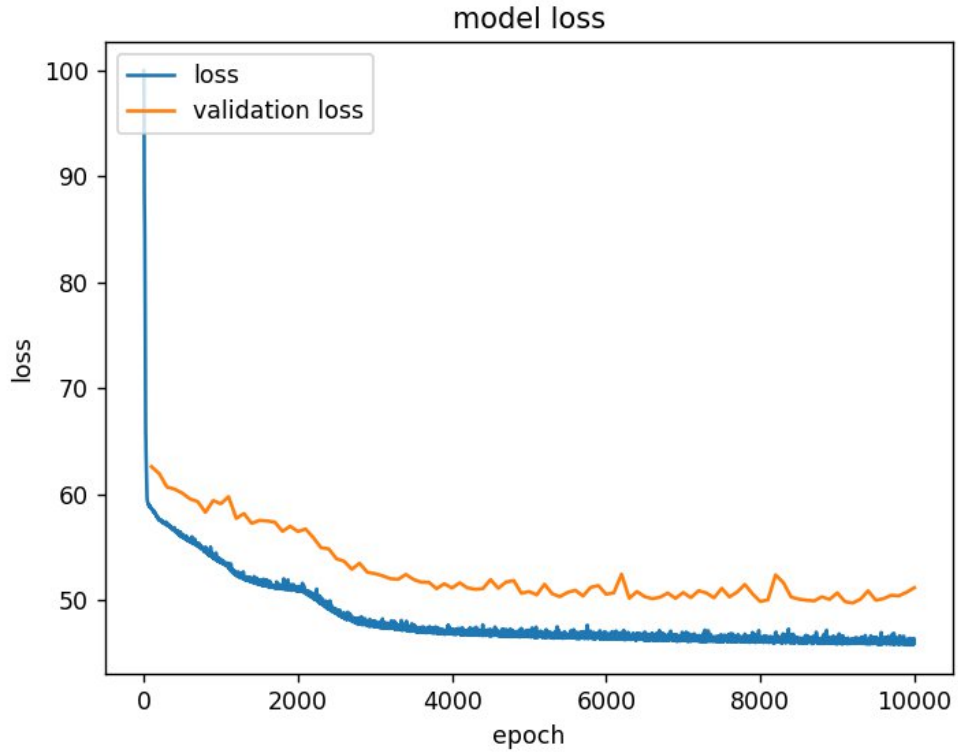

Рис. 2.5.(1) Четверта ітерація моделі

Результат у середньому дає передбачення з похибкою 44%.

Валідація на основі даних з бази для тестування дає похибку 46%. Це означає, що модель не зазнала перенавчання і похибка моделі відповідає похибці на основі даних, які не використовувались для тренування (дані які модель ще не бачила). З цього можна зробити висновок, що модель має перспективи на покращення в разі, якщо знайти кращу конфігурацію, зробити

повторну оцінку вагу важливості кожного окремого параметру, або збільшення часу тренування.

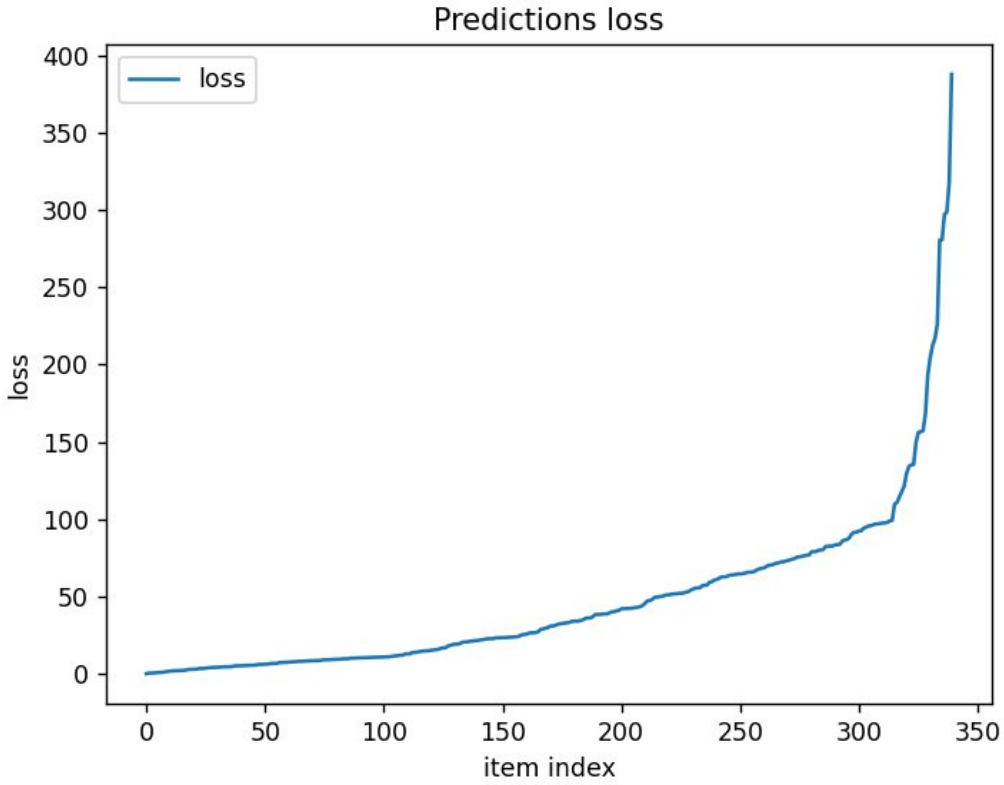

Рис. 2.5.(2) Графік похибок

Графік показує відсортовані похибки кожного окремого елементу з бази для тестування. Так, як було доведено відповідність похибок моделі під час тренування та валідації на базі тестових даних. Наявність незначної кількості похибок значно більших за середню може означати або некоректно визначену ціну об’єкту, або те, що ціна об’єкта була визначена за параметрами не відображеними у даній базі даних (соціоекономічні фактори).

Висновки до розділу 2

У цьому розділі була проведена реалізація методів аналізу вартості земельних ділянок за супутниковими та геопросторовими даними. Починаючи з вибору датасету і відповідних даних, до досягнення кращих результатів

шляхом ітерацій моделі, робота зосереджувалася на важливих аспектах аналізу.

Було проведено детальний аналіз датасету. Вибір датасету був здійснений на основі вимог проекту та доступних даних. Були відібрані певні параметри, які вважалися релевантними для аналізу вартості земельних ділянок. Поділ датасету на тренувальні, валідаційні та тестувальні набори був здійснений для забезпечення ефективного навчання та оцінки моделі. Нормалізація даних була виконана для стандартизації значень та забезпечення стабільності обчислень. Аналіз кореляції даних допоміг виявити зв'язки між параметрами та оцінити їх вплив на вартість земельних ділянок.

Було використано один внутрішній шар та функцію активації SoftPlus. Хоча модель показала певний рівень точності в передбаченні вартості земельних ділянок, виявилася обмеженою у здатності передбачати ціну з високою точністю.

Далі, було внесено покращення до моделі. Було додано два внутрішні шари та функцію активації ReLU. Це призвело до покращення точності моделі, проте результати залишилися нестабільними та хаотичними.

Було спробувано використати метод Random Forest для аналізу вартості земельних ділянок. Цей метод показав найгірші результати. Random Forest забезпечив більшу стабільність та узагальнюючу здатність моделі завдяки використанню дерев рішень, однак дуже погану точність.

В останній моделі було досягнуто найменшої можливої похибки, та була доведена відповідність похибок моделі під час тренування та похибок на основі бази даних яка не використовувалась для процесу тренування. Була виявлена відносно невелика кількість даних, що мали аномально велику похибку. Відносно невелика точність підтверджує теорію про складність роботи в оцінці земельних ділянок за супутниковими даними (геопросторовими) через низьку кореляцію з деякими проведенеми оцінками. Однак об’єктивні оцінки надані саме за вищевказаними даними

підтверджують корисність моделі у оцінці великої кількості різних ділянок, по яким ще не була проведена ЕГО.

Загалом, реалізація методів аналізу вартості земельних ділянок за супутниковими та геопросторовими даними показала значний прогрес. Проте, можливо, існує потенціал для подальшого вдосконалення моделі шляхом впровадження додаткових методів та оптимізації параметрів.

ВИСНОВКИ

В даній дипломній роботі було проведено дослідження методів аналізу вартості земельних ділянок з використанням супутникових та геопросторових даних. У першому розділі були розглянуті теоретичні аспекти, пов'язані з оцінкою земельної вартості, включаючи різні методи оцінки, такі як нормативна грошова оцінка та експертна грошова оцінка. Також була розглянута актуальність використання супутникових даних у сфері аналізу вартості земельних ділянок, оскільки вони надають значну кількість інформації про земельну площу та місцевість.

У другому розділі була проведена реалізація методів аналізу вартості земельних ділянок за супутниковими та геопросторовими даними. Було виконано детальний аналіз датасету, включаючи вибір та обробку даних, нормалізацію та аналіз кореляції. Також було розглянуто вибір та конфігурацію моделі, включаючи різні типи регресій та функції активації.

Під час експериментів з різними моделями, включаючи нейронні мережі та метод Random Forest, було виявлено, що оцінка вартості земельних ділянок за допомогою супутникових та геопросторових даних є складним завданням. Моделі показали обмежену здатність передбачати ціну земельних ділянок з високою точністю, особливо через низьку кореляцію з деякими проведеними оцінками.

При використанні супутникових та геопросторових даних модель може допомогти у визначенні загальних тенденцій і розумінні факторів, що впливають на вартість земельних ділянок. Вона може бути корисною для роботи з великими обсягами даних і виявленням регіональних або місцевих змін у цінах на землю. Однак, важливо пам'ятати, що результати моделі потрібно перевіряти і порівнювати з реальними транзакціями та експертними оцінками, оскільки є багато інших факторів, що можуть впливати на вартість земельних ділянок, таких як інфраструктура, доступність до міста, забудованість та інші.

Загалом, використання супутникових та геопросторових даних для аналізу вартості земельних ділянок відкриває нові можливості для геодезистів, економістів та інших спеціалістів. Дослідження у цій області може сприяти покращенню точності оцінки вартості землі, виявленню ринкових тенденцій та плануванню місцевого розвитку. Однак, потрібно продовжувати дослідження та розвивати нові методи та моделі для отримання більш точних прогнозів вартості земельних ділянок.

ПЕРЕЛІК ПОСИЛАНЬ

1) К. М. Ніколайчук, В. С. Мошинський, А. Г. Ліщинський “Методичні вказівки до виконання самостійної роботи з дисципліни “Оцінка землі та нерухомого майна””, 2013
2) Jia X. and Weng Q. Mapping land values using big geospatial data and machine learning. International Journal of Geographical Information Science, 33(6), 1191-1212, 2019
3) Bin, S., & Liu, X. Land value estimation using machine learning techniques: A review. ISPRS International Journal of Geo-Information, 9(3), 155, 2020
4) Czepiel J. & Kaczmarek T. Using remote sensing data for land valuation purposes. Polish Journal of Environmental Studies, 24(6), 2299-2307, 2015
5) Huang S., Yu J., Yin H., & Wang H. Determining the market value of rural residential land: a case study in Beijing. Sustainability, 11(9), 2487, 2019
6) Ahmed D., Yuji M., A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan, 2020
7) https://www.fastaireference.com/overfitting
8) https://medium.com/analytics-vidhya/linear-regression-and-polynomial-regression-using-normal-equation-method-c3929d71734d
9) https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/
10) https://pytorch.org/docs/stable/generated/torch.nn.PReLU.html
11) https://hvidberrrg.github.io/deep_learning/activation_functions/sigmoid_function_and_derivative.html
12) https://www.kenhub.com/en/library/anatomy/action-potential

ДОДАТОК А. ТЕКСТ ПРОГРАМИ

А.1 Програма 1. Фінальна ітерація моделі

```
import os, sys, csv, shutil, time
import tensorflow as tf
from sklearn.model_selection import train_test_split
from matplotlib import pyplot as plt

# Стовпці після конвертації
COLUMNS = ['state_region', 'cadastral', 'ownership', 'land_category', 'purpose_number',
'purpose_description', 'fraction_s', 'right_name', 'cd_name', 'oblast', 'ftype', 'area', 'right_type',
'effective', 'cost_of_ri', 'cd_type', 'term_year', 'priceha', 'ppoint', 'lc_2016', 'lc_2017', 'lc_2018',
'lc_2019', 'lc_2020', 'lc_2021', 'bonitet', 'ndvi_2019', 'ndvi_2020', 'ndvi_2021', 'altitude', 'slope',
'dist_citie', 'dist_towns', 'dist_villa', 'dist_eleva', 'dist_pr_ro', 'dist_sec_r', 'dist_count', 'bonitetl',
'kyiv_dist', 'mjr_cr', 'area_mon', 'purpose', 'deal_type', 'price', 'right_registration_date',
'land_evaluation_norm_price', 'evaluation_date', 'count_price_ha']

# Дані для тренування моделі
TRAINING_DATA_COLUMNS = ['area', 'ndvi_2021', 'kyiv_dist', 'dist_citie', 'dist_towns',
'dist_villa', 'dist_eleva', 'dist_pr_ro', 'dist_sec_r', 'dist_count', 'bonitetl', 'mjr_cr']

# Вивід моделі
TRAINING_DATA_TARGET = 'price'

# Дані для нормалізації вводу
DISTANCE_COLUMNS = [    (COLUMNS.index('kyiv_dist'),          1000,
        10.535763775011948),
                        (COLUMNS.index('dist_citie'),         800,    0.011929692467384958),
                        (COLUMNS.index('dist_towns'),         600,    0.015061359623583915),
                        (COLUMNS.index('dist_villa'),         300,    0.1370802573740055),
                        (COLUMNS.index('dist_eleva'),         200,    0.05435184831945874),
                        (COLUMNS.index('dist_pr_ro'),         500,    0.02228859346255508),
                        (COLUMNS.index('dist_sec_r'),         200,    0.012363372823044397),
                        (COLUMNS.index('dist_count'),         200,    0.0638229421839469)
                    ]       # (idx, a, mult), y=a/(x+a) * mult
OTHER_COLUMNS =[        (COLUMNS.index('area'),               1),
                        (COLUMNS.index('ndvi_2021'),        1. / 9 * 0.08943301695218853),
```

```
                    (COLUMNS.index('bonitetl'),   1. / 80 * 0.03914624078350417),
                    (COLUMNS.index('mjr_cr'),     1. / 71 * 0.01066955571236979),
                    (COLUMNS.index('lc_2021'),    1. / 18 * 0.00808934528600921)
                ]           # others multipliers
################

def load_db(path):
        # Читаємо базу даних
        with open(path, 'r', encoding = 'utf-8') as file:
                table = csv.reader(file, delimiter=';', quotechar="")
                data = [i for i in table]
        return data

def row_ha_convert(row):
        # Конвертуємо площу у га
        area = float(row[11]) / 10000 if row[12] == 'кв.м' else float(row[11])

        return row[:11] + [area] + row[13:]

def row_convert(row):
        # Data conversions
        # Distances to one number
        distance_coefs = [ a / (float(row[idx]) + a) * mult for idx, a, mult in
DISTANCE_COLUMNS]
        # End conversions
        return [float(row[idx]) * mult for idx, mult in OTHER_COLUMNS] + distance_coefs



def prepare_training_data(raw_data):
        data = [row_ha_convert(i) for i in raw_data]
        # Масив індексів даних для вводу і результату
        outp_idx = COLUMNS.index(TRAINING_DATA_TARGET)
        input_data = []
        output_data = []
        no_result_data = []
        errors = []
        for row in data:
                try:
```

```
                    inp_convert = row_convert(row)

                    if row[outp_idx] == '':
                        no_result_data.append(inp_convert)
                        continue

                    # Price for 1 ha
                    outp = float(row[outp_idx])

                    input_data.append(inp_convert)
                    output_data.append([outp])

                except ValueError as e:
                    errors.append(str(e))

                except Exception as e:
                    errors.append(str(e))


        return input_data, output_data, no_result_data, errors

def prepare_predict_data(data, cads):
        cad_idx = data[0].index('cadastral')
        output = []
        cads_output = []
        for row in data[1:]:
            if row[cad_idx] in cads:
                cadastral = row[cad_idx]
                cads.remove(cadastral)
                try:
                    output.append(row_convert(row_ha_convert(row)))
                    cads_output.append(cadastral)
                except:
                    continue
        return output, cads_output, cads

def backup_name(name):
        idx = name.rfind('.')
```

```
        backup = '_backup_%s' % time.strftime('%y-%m-%d_%H:%M:%S')
        return name[:idx] + backup + name[idx:]

def get_model(name, size=len(TRAINING_DATA_COLUMNS)):
        # Завантажуємо модель або створюємо нову
        if os.path.exists(name):
                model = tf.keras.models.load_model(name)
        else:
                model = tf.keras.Sequential([
                tf.keras.layers.Dense(64, activation='relu'),
                tf.keras.layers.Dense(64, activation='relu'),
                tf.keras.layers.Dense(1)
                ])
                model.compile(loss='mean_absolute_percentage_error',
                optimizer=tf.keras.optimizers.Adam(learning_rate=0.01))
        return model

def save_model(model, name):
        if os.path.exists(name):
                shutil.copyfile(name, backup_name(name))
        model.save(name)

match sys.argv[1]:
        case 'train':
                # Тренування моделі
                name = 'models/' + sys.argv[2] + '.h5'
                epochs = int(sys.argv[3])
                validation_freq = 100

                i, o, n, e = prepare_training_data(load_db('train_data.csv')[1:])

                tr_X, val_X, tr_Y, val_Y = train_test_split(i, o, test_size=0.1)
                model = get_model(name)
                history = model.fit(tr_X, tr_Y, validation_data=(val_X, val_Y),
validation_freq=validation_freq, epochs=epochs, steps_per_epoch=10, verbose=2)
                save_model(model, name)

                plt.plot(history.history['loss'])
```

```
                plt.plot(range(validation_freq, epochs + 1, validation_freq),
history.history['val_loss'])
                plt.title('model loss')
                plt.ylabel('loss')
                plt.xlabel('epoch')
                plt.legend(['loss', 'validation loss'], loc='upper left')
                plt.show()

        case 'validate':
                # Перевірка похибки моделі
                name = 'models/' + sys.argv[2] + '.h5'
                i, o, n, e = prepare_training_data(load_db('test_data.csv')[1:])
                model = get_model(name)
                res = model.evaluate(i, o, verbose=1)
                print(res)

        case 'predict':
                name = 'models/' + sys.argv[2] + '.h5'
                cads = sys.argv[3:]

                data = load_db('data.csv')
                inp, cads, errors = prepare_predict_data(data, cads)
                if len(inp) == 0:
                        print("Не знайдено жодного кадастрового номеру!")
                        exit(1)

                model = get_model(name)
                out = model.predict(inp, verbose=1)

                print("Ціни:")
                print(*['%25s : %12.2f' % (cads[i], inp[i][0]) for i in range(len(out))])

                if len(errors) > 0:
                        print("Не знайдено ціни для:")
                        print(*errors, sep='\n')

        case 'predict_validate':
                name = 'models/' + sys.argv[2] + '.h5'
```

```
                i, o, n, e = prepare_training_data(load_db('test_data.csv')[1:])

                model = get_model(name)
                out = model.predict(i, verbose=1)
                stats = [(o[idx][0], out[idx][0], abs(o[idx][0] - out[idx][0]) / o[idx][0] * 100) for
idx in range(len(i))]

                print(*['%10.2f %10.2f     %10.2f' % stat for stat in stats], sep='\n')

                plt.plot(sorted([i[2] for i in stats]))
                plt.title('Predictions loss')
                plt.ylabel('loss')
                plt.xlabel('item index')
                plt.legend(['loss'], loc='upper left')
                plt.show()

        case 'split_db':
                db = load_db('data.csv')
                t_idx = db[0].index(TRAINING_DATA_TARGET)
                c_idx = db[0].index('purpose_number')
                db = [i for i in db if i[t_idx] != '' and i[c_idx] not in ['', 'NULL']]

                train, test = train_test_split(db[1:], test_size=0.02)

                with open('train_data.csv', 'w', encoding='utf-8', newline='') as file:
                        writer = csv.writer(file, delimiter=';')
                        writer.writerows([db[0]] + train)
                with open('test_data.csv', 'w', encoding='utf-8', newline='') as file:
                        writer = csv.writer(file, delimiter=';')
                        writer.writerows([db[0]] + test)
        case 'test':
                data = load_db('data.csv')
                data = data[1:]
                d = {i : 0 for i in LAND_CLASSES}
                for i in data:
                        d[i[4]] += 1
                print(d)
```