**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ”**

**ФАКУЛЬТЕТ ІНФОРМАТИКИ**

**КАФЕДРА МАТЕМАТИКИ**

**ЗВІТ**

**з дослідницької практики на тему:**

**“Random forest”**

Студента 3-го курсу

спеціальності 113

“Прикладна математика”

Вітіска Владислава Олександровича

**Київ – 2024**

**Анотація**

Практика присвячена дослідженню методу машинного навчання Random forest.

Вперше ансамблевий алгоритм на основі дерев прийняття рішень, що носив назву «Random decision forests», був розроблений та описаний вченим IBM Тіном Кам Хо у 1995, після чого алгоритм був покращений Лео Брейманом і Аделем Катлером, що і дали йому звичну назву «Random Forest», яку запатентували. Сьогодні випадковий ліс – один з найпопулярніших методів як для задач класифікації, так і для регресії.

У даній роботі розглядається Random forest, його теоретична та практична сторона. Протягом практики було розглянуто створення моделей на основі набору даних для передбачення ядовитості грибу за його зовнішніми характеристиками, набору для передбачення смерті спричиненої серецевою недостатністю та набору для передбачення цінової категорії телефону на основі його технічних характеристик. Прикладом для подібної роботи та матеріалом з більш загальним поясненням роботи методу була робота Md Atikur Rahman, Akihiko Hokugo Nobuhito Ohtsu “Household evacuation preparation time during a cyclone: Random Forest algorithm and variable degree analysis” (2021). Також оглядово було розглянуто інші методи машинного навчання та порівняння результатів з алгоритмом випадкового лісу.

У першому розділі звіту розглядаються основні поняття та принципи роботи дерев прийняття рішень. У другому розділі та його підпунктах досліджуються принципи, що застосовуються для бінарних дерев для утворення методу випадкового лісу. В третьому розділі було створено різні моделі для дослідження якості передбачення Random Forest та впливу гіперпараметрів на вихідні передбачення. В четвертому розділі на основі понять та коду використаних в третій главі було проведено дослідження ще двох наборів даних для можливості більш широкого порівняння з іншими алгоритмами та впливу гіперпараметрів на отримані результати.

**Розділ 1. Дерева прийняття рішень**

**Означення 1.1.** Деревом називається граф G = (V, E), у якому довільні дві вершини (або вузли) з'єднані рівно одним шляхом.

**Означення 1.2.** Деревом з коренем називається дерево, в якому одна з вершин позначена як корінь. У нашому випадку ми додатково припускаємо, що кореневе дерево є орієнтованим графом, де всі ребра спрямовані від кореня.

**Означення 1.3.** Якщо існує ребро з t1 в t2 (тобто, якщо (t1, t2) ∈ E), то кажуть, що вершина t1 є батьком вершини t2, а вершина t2 є нащадком вершини t1.

**Означення 1.4.** У кореневому дереві вершина називається внутрішньою, якщо вона має одного або більше нащадків, і термінальною, якщо вона не має нащадків. Термінальні вершини також називають листям.

**Означення 1.5.** Бінарним деревом називається кореневе дерево, у якого усі внутрішні вершини мають рівно двох нащадків.

Використовуючи вище наведені означення, ми можемо визначити дерево прийняття рішень, як модель , що представлена кореневим деревом (часто бінарним, але не обов'язково), де будь-яка вершина цього дерева t представляє підпростір  вхідного простору, з кореневою вершиною , що відповідає самому . Внутрішні вершини позначено поділом , взятим з набору запитань Q. Вона розбиває простір , який представляє вершина , на непересічні підпростори що відповідають кожному з його нащадків.

**Наприклад.**

Розглянемо задачу бінарної класифікації для точки з координатами

, де точку віднесемо до класу c1 якщо вона задовольняє умову

, інакше це буде точка класу c2.

Розглянемо дерево прийняття рішень (рис. 1.1)

Зображення, що містить схема, ряд, коло

Автоматично згенерований опис

**рис. 1.1**

На рисунку маємо корінь , внутрішню вершину , та три термінальні вершини: , , .

Не термінальні вершини та мають поділи:

* -
* -

Якщо розглядати утворені вершинами підрозділ, то наглядно це можна зрозуміти з рис. 1.2

Зображення, що містить Барвистість, схема, Прямокутник, візерунок

Автоматично згенерований опис

**рис. 1.2**

У вершині нашим простором буде весь прямокутник, у вершині – підпростір обмежений червоним прямокутником, – обмежений зеленим прямокутником, після підпростір поділиться на дві частини, одна з яких обведена синім, що містить точки класу с1.

Оскільки ми розглянули простий приклад задачі з класифікації, варто сказати, що для термінальних вершин в задачах регресії результатом передбачення , що прив’язаний до відповідної термінальної вершини, буде не клас, як для задач класифікації, а дійсне число, яке буде найкращим припущенням.

Формально задача передбачення за допомогою дерева прийняття рішень може бути описана наступним алгоритмом-псевдокодом

**Алгоритм 1.1** Отримання результату передбачення в дереві прийняття рішень

1: **Функція** Передбачити()

2:

3: **поки** не є термінальною вершиною виконувати

4: вершина нащадок , така що

5: **закінчення блоку**

6: **повернути**

7: **закінчення функції**

Отже, ми розібрались як за допомогою дерева рішень зробити прогноз, але очевидно, що для задач машинного навчання подібне дерево не відоме одразу і є ціллю розробки моделі.

Суть побудови дерева прийняття рішень є створення моделі, що найкраще підходитиме для відомого тренувального набору , причому серед усіх дерев прийняття рішень можуть існувати кілька найкращих. В таких випадках згідно з принципом Бритви Оккама, обирається варіант з найменшою кількістю припущень, що в нашому випадку полягає в меншому розмірі дерева, або ж це те ж саме що і менша кількість внутрішніх вершин.

Знаходження найменшого дерева є NP-повною задачею. Як наслідок, не існує не існує ефективного алгоритму для знаходження , тим самим припускаючи, що знаходження близьких до оптимальних дерев рішень є найкращим рішенням, щоб утримати вимоги до обчислень у реалістичних межах реалістичності.

Введемо загальне поняття забрудненості. Визначимо міру забрудненості i(t) як функцію, яка оцінює доброякісність будь-якої вершини t. Припустимо, що чим менше i(t), тим чистіша вершина і тим краще передбачення для усіх , де – підмножина навчальних прикладів що потрапляють до t. Починаючи з однієї вершини, наближене до оптимального дерева може бути отримане жадібним ітеративним методом розподіляючи вершини на кілька чистіших, до поки існують внутрішні вершини, які можна розбити на чистіші.

Жадібне припущення про те, що отримане дерево має бути якомога меншим, тобто мати гарне узагальнення, полягає в тому, щоб розбити кожну вершину t за допомогою розбиття s, яке локально максимізує зменшення забруднення у отриманих дочірніх вершинах.

Для формалізації введемо наступне означення

Означення 3.6 Зменшення забруднення бінарного розподілу , що ділить вершину t на праву та ліву вершини це

Де або – відношення кількості прикладів, що потрапили до дочірньої вершини або , до кількості прикладів що були у батьківській t.

Використовуючи це означення ми можемо записати алгоритм для побудови дерева прийняття рішень з бінарними розподілами .

**Алгоритм 2.** Жадібна побудова бінарного дерева рішень.

1: **Функція** ПобудуватиРішенняДерева(L)

2: **Створити** дерево рішень з кореневим вузлом

3: **Створити** порожній стек S відкритих вузлів (t,)

4: S.push((,L))

5: **поки** S не порожній робити

6: t, = S.pop()

7: **якщо** для t виконується критерій зупинки тоді

8: = деяке постійне значення

9: **інакше**

10: Знайти розділення на , яке максимізує

s\* = arg max s∈Q Δi(s, t)

11: Розділити на і відповідно до s\*

12: Створити лівий дочірній вузол для t

13: Створити правий дочірній вузол для t

14: S.push((,))

15: S.push((,))

16: **кінець якщо**

17: **кінець поки**

18: **повернути**

19: **кінець функції**

Схожим чином можна реалізувати і алгоритм побудови дерева з N-арним розподілом, втім бінарного розподілу достатньо для подальшого розуміння в контексті методу випадкового лісу.

**Розділ 2. Random Forest**

У деревах прийняття рішень є значна проблема, що понижує якість вихідного передбачення: висока дисперсія. Втім цей метод має низький зсув.

**Означення 2.1** Зсув (англ. bias) — це похибка, викликана помилковими припущеннями в алгоритмі навчання.

Розумний підхід до зменшення помилки прогнозування полягає у зменшенні дисперсії прогнозу за умови, що відповідний зсув може залишатися незмінним або не збільшуватися надто сильно. Рішенням цієї проблеми є ансамблеві методи. Зокрема, основний принцип ансамблевих методів, заснованих на рандомізації, полягає у введенні випадкових збурень у процедуру навчання, щоб створити кілька різних моделей з однієї навчальної вибірки L, а потім об'єднати прогнози цих моделей для формування прогнозу ансамблю.

Random forest – це сімейство ансамблевих методів машинного навчання, що використовує численні дерева прийняття рішень для створення прогнозів.

Розглянемо два ключові ансамблеві мета-алгоритми, що використовуються в Random Forest – одній з найпопулярнішиї варіацій сімейства відповідних методів, розробленій Лео Брейманом і Аделем Катлером.

**2.1 Метод Bagging або ж Bootstrap aggregating.**

Суть Бутстрепової агрегації полягає у спеціальній вибірці набору даних з L для тренування дерева: з відповідного початкового тренувального набору розміру береться елементів з можливими повторами. Для випадку в середньому майже 37% тренувальних прикладів не потраплять до нового набору .

**2.2 Метод випадкових підпросторів.**

Дослідник Тін Кам Хо запропонував при відборі даних для тренування нових дерев звертати увагу на відбір атрибутів, оскільки важливі параметри можуть привести до кореляції значної кількості дерев, що потенційно може вплинути на якість передбачень. Для задач класифікації зазвичай використовується можливих параметрів, а для регресії .

**2.3 Прийняття рішення.**

Після утворення дерев прийняття рішення для утворення єдиного передбачення для задач регресії та класифікації використовуються різні методи: у першому випадку використовується середнє значення отриманих результатів, а в другому – обирається клас з найбільшою кількістю голосів.

Узагальнюючи, метод випадкових лісів для задачі класифікації можна зобразити наступним чином(рис. 2.1)

Зображення, що містить текст, схема, знімок екрана, План

Автоматично згенерований опис

**рис. 2.1**

**Розділ 3. Практичне дослідження методу випадкових лісів.**

**3.1 Знайомствто з основним набором даних.**

Оскільки ціллю дослідження є вивчення методу випадкового лісу було обрано набір даних з задачею класифікації чи ядовитий гриб. Цей набір – очищений від порожніх значень та оброблений щоб перевести категоріальні буквенні змінні до числових.

Маємо 9 колонок:

* Розмір шапки гриба
* Форма шапки
* Вид кріплення пластинки гриба
* Колір пластинки
* Висота ніжки
* Товщина ніжки
* Колір ніжки
* Пора року
* Чи ядовитий гриб – цільова колонка

До набору входить 54035 записів, а кожна колонка має кількість унікальних значень зображену на рис. 3.1 з розподілом вказаним на рис 3.2

Зображення, що містить текст, знімок екрана, число, Шрифт

Автоматично згенерований опис

**рис. 3.1**

Зображення, що містить знімок екрана

Автоматично згенерований опис

**рис. 3.2**

Варто уточнити, що під час обробки інформації автор датасету перетворив пори року на число з комою, хоча унікальних значень залишилось 4.

**3.2 Правила Тренування Моделі**

Процес тренування буде відбуватися за допомогою бібліотеки scikit-learn в середовищі Jupiter Notebook на мові програмування Python.

Під час роботи над моделями було висунуте припущення, що стандартні параметри моделі задані розробниками бібліотеки є оптимальними, і порівняння відбувається базових варіантів, за виключенням одного методу, про що буде написано в частині порівняння різних алгоритмів.

Для відтворюваності дослідження, оскільки метод випадкового лісу, а саме алгоритми Bootstrap Aggregating та Метод Випадкових підпросторів, ґрунтується на випадкових процесах, було згенероване значення 215 з діапазону [0, 1000], яке буде передаватись моделі як основа для випадкової генерації.

Для тренування відбиралось 80% даних, решта 20% використовувалась для тестування моделі. Для відтворення результатів було використане значення 0 для випадкової генерації розподілу.

**3.2.1 Методи порівняння результатів передбачення.**

Як критерії оцінювання якості було обрано три статистики

1. Точність – ступінь наближення ряду передбачень до справжніх значень.

Для задач класифікації в загальному вигляді можна записати наступну формулу:

1. Влучність – ступінь наближення вимірювань один до одного.

Для задач бінарної класифікації описується наступним чином:

1. F1-міра – середнє гармонійне влучності та повноти

Формула:

**3.3 Дослідження випадкового лісу**

**3.3.1 Тренування базової моделі**

Після тренування моделі були отримані наступні результати:

Точність - 0.9907467382252244;

Влучність - 0.9907376220949815;

F1-міра - 0.9915725602561941.

Ці результати будуть використані для подальших порівняннь.

**3.3.2 Дослідження впливу гіперпараметрів**

Означення 3.1 Гіперпараметр — це параметр, значення якого використовується для керування процесом навчання.

В наступних пунктах буде досліджено вплив кількості дерев прийняття рішень, вплив використання Bootstrap Aggregating, вплив кількості значень при виборі випадкового простору та вплив максимальної глибини дерева.

**3.3.3 Вплив кількості дерев**

Базовим значенням для реалізації класифікатора, що використовує випадковий ліс, з бібліотеки scikit-learn є 100 дерев. Розглянемо, як вплине зміна кількості дерев на точність передбачення. Було обрано перевірити випадки з 1, 2, 5, 10, 25, 50, 100, 200, 500 деревами. Результати цього(рис 3.3) та подальших порівнянь будуть у вигляді HeatMap створених за допомогою бібліотеки Seaborn.

Зображення, що містить текст, знімок екрана, число, Шрифт

Автоматично згенерований опис

**рис. 3.3**

Як бачимо, значна просадка якості передбачень є тільки для 1 та 2 дерев.

**3.3.4 Вплив Bootstrap Aggregating**

Як було описано в теоретичній частині, цей алгоритм допомагає підвищити точність передбачень. Незначне зниження точності можна побачити графічно на рис 3.4, втім різниця вимірюється в тисячних, що не сильно впиває на якість отриманих результатів. Протягом усіх подібних досліджень цього методу на цьому наборі даних різниця буде мала, причини чого ми розглянемо пізніше.

Зображення, що містить текст, знімок екрана, Шрифт, дизайн

Автоматично згенерований опис

**рис. 3.3**

**3.3.5 Вплив числа параметрів**

Другим важливим алгоритмом, що є складовою випадкових лісів, є метод випадкових просторів. Теорія каже нам для задач класифікації використати корінь з числа параметрів значень, що є значенням за замовчуванням в цій бібліотеці. На рис 3.4 можна побачити результати для 1, 6, 8 та фактично 2, оскільки кількість параметрів – 8, а ціла частина від кореня буде 2. Корінь показав найкращий резульатат, але знову різниця в тисячних.

Зображення, що містить текст, знімок екрана, Шрифт, число

Автоматично згенерований опис

**рис. 3.4**

3.3.5 Влив максимальної глибини дерева

Одним з гіперпараметрів доступних для модифікації є максимальна глибина дерева – найбільша можлива відстань від кореня дерева до найвіддаленішого нащадка. На рис 3.5 можна побачити вплив максимальної глибини заданої значеннями 1, 2, 5, 10, 25, 50, 100, 200, 500. Варто сказати, що за замовчуванням обмежень на цей параметр немає. Як бачимо після 25 точність незмінна.

Зображення, що містить текст, знімок екрана, число, Шрифт

Автоматично згенерований опис

**рис. 3.5**

**3.4 Порівняння методу випадкового лісу з іншими алгоритмами**

Для порівняння було відібрано 5 інших методів класифікації, серед яких:

* Дерево прийняття рішень;
* Логістична регресія;
* Метод опорних векторів;
* Наївний баєсів класифікатор;
* Метод k-найближчих сусідів.

Відбір алгоритмів був здійснений на різних статтях, щодо популярних методів класифікації, а дерево прийняття рішень буде корисним джерелом інформації про вплив покращень застосованих в «випадкових лісах». Всі методи були натреновані зі стандартними параметрами, окрім логістичної регресії, для якої було збільшено кількість можливих ітерацій до 10000.

Дивлячись на результати(рис 3.6), можна помітити, що зі значним відривом найкращими моделями є дерево прийняття рішень та Random forest, що пояснює низький вплив зміни гіперпараметрів, оскільки алгоритм на якому ґрунтується випадковий ліс дає гарні передбачення, а покращений метод хоч і має кращі передбачення, втім простір для покращень досить обмежений. Решта методів дали схожі результати, що значно відстають від перших.

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

**рис. 3.6**

**Розділ 4. Додаткові дослідження**

Отримавши результати з розділу 3 та маючи весь необхідний код, для перевірки та порівняння якості передбачень було проведено дослідження на базі ще двох наборів даних: задача бінарної класифікації для передбачення ризику смерті пацієнта, пов’язаної з серцевою недостатністю, та задача поділу телефонів на 4 цінові категорії на основі технічних характеристик пристрою.

Принципи дослідження залишились незмінними як і код, за виключенням видалення логістичної регресії для задачі розподілу на 4 категорії, оскільки вказаний метод розроблений для роботи з бінарною класифікацією, та методу опорних векторів для задачі передбачення проблем з серцем через помилки в оцінюванні якості передбачення.

У даному розділі для обох наборів даних буде проведено лиш поверхневий аналіз отриманих результатів.

**4.1 Передбачення серцевої недостатності**

Маємо набір даних з 5000 рядків, що має наступні атрибути:

* age: вік пацієнта (роки)
* anaemia: зменшення кількості еритроцитів або гемоглобіну (булевий)
* creatinine phosphokinase (CPK): рівень ферменту КФК у крові (мкг/л)
* diabetes: чи є у пацієнта цукровий діабет (булевий)
* ejection fraction: відсоток крові, що виходить з серця при кожному скороченні (у відсотках)
* high blood pressure: чи є у пацієнта гіпертонія (булеве значення)
* platelets: кількість тромбоцитів у крові (кілотромбоцитів/мл)
* sex: жінка чи чоловік ( бінарний тип)
* serum creatinine: рівень креатиніну в крові (мг/дл)
* serum sodium: рівень натрію в крові (мекв/л)
* smoking: курить пацієнт чи ні (булеве)
* time: період спостереження (дні)
* DEATH\_EVENT: чи помер пацієнт протягом періоду спостереження (булеве). Цільова колонка.

Після простого тренування Random Forest була отримана наступна точність:

Точність - 0.992;

Влучність - 0.9904761904761905;

F1-міра - 0.9873417721518988.

Результати тестування різної кількості дерев показані на рис. 4.1.1, впливу Bootstrap Aggregating на рис 4.1.2, кількості значень при виборі випадкового простору на рис 4.1.3 та вплив максимальної глибини дерева(рис 4.1.4). На рис. 4.1.5 зображено порівняння різних алгоритмів.

Зображення, що містить текст, знімок екрана, число

Автоматично згенерований опис

**рис. 4.1.1**

Зображення, що містить текст, знімок екрана, дизайн

Автоматично згенерований опис

**рис. 4.1.2**

Зображення, що містить текст, знімок екрана, дизайн

Автоматично згенерований опис

**рис. 4.1.3**

Зображення, що містить текст, знімок екрана, число, Шрифт

Автоматично згенерований опис

**рис. 4.1.4**

Зображення, що містить текст, знімок екрана, схема, дизайн

Автоматично згенерований опис

**рис. 4.1.5**

Загалом, результати дослідження цього набору даних схожі до основного дослідження.

**4.2 Передбачення цінової категорії телефону**

Маємо набір даних з 2000 рядків, що має наступні атрибути:

* battery\_power: максимальна ємність батареї в мАг
* blue: Має bluetooth чи ні
* clock\_speed: частота процесора
* dual\_sim: Має підтримку двох SIM-карт чи ні
* fc: Фронтальна камера в мегапікселях
* four\_g: Має 4G чи ні
* int\_memory: Внутрішня пам'ять у гігабайтах
* m\_dep: Глибина мобільного в см
* mobile\_wt: Вага мобільного телефону
* n\_cores: Кількість ядер процесора
* pc: Основна камера, мегапікселів
* px\_height: Висота роздільної здатності у пікселях
* px\_width: Ширина роздільної здатності пікселів
* ram: Оперативна пам'ять у мегабайтах
* sc\_h: Висота екрану мобільного в см
* sc\_w: Ширина екрану мобільного в см
* talk\_time: максимальний час розмови від одного заряду батареї.
* three\_g: Є 3G чи ні
* touch\_screen: Є сенсорний екран чи ні
* wifi: є wifi чи ні
* price\_range: Це цільова змінна зі значеннями 0 (низька ціна), 1 (середня ціна), 2 (висока ціна) та 3 (дуже висока ціна).

Після тренування Random Forest без змін гіперпараметрів були отримані такі результати якості передбачення:

Точність - 0.855;

Влучність - 0.8500110537951364;

F1-міра - 0.850575886471355.

Результати тестування різної кількості дерев показані на рис. 4.2.1, впливу Bootstrap Aggregating на рис 4.2.2, кількості значень при виборі випадкового простору на рис 4.2.3 та вплив максимальної глибини дерева(рис 4.2.4). На рис. 4.2.5 зображено порівняння різних алгоритмів.

Зображення, що містить текст, знімок екрана, число

Автоматично згенерований опис

**рис. 4.2.1**

Зображення, що містить текст, знімок екрана

Автоматично згенерований описЗображення, що містить текст, знімок екрана, схема, число

Автоматично згенерований опис

**рис. 4.2.3**

**рис. 4.2.2**

Зображення, що містить текст, знімок екрана, число, Шрифт

Автоматично згенерований опис

**рис. 4.2.4**

Зображення, що містить текст, знімок екрана

Автоматично згенерований опис

**рис. 4.2.5**

В цьому дослідженні є більш виразні зміни в результатах налаштування гіперпараметів, але для цього набору даних метод випадкових лісів поступився в якості передбачення іншим.

Висновки

Під час практики було досліджено метод Random Forest.

В першому розділі роботи були розглянуті основні поняття пов’язані з деревами прийняття рішень, необхідними для розуміння досліджуваного методу.

В другому розділі були розглянуті основні алгоритми застосовані в методі Random Forest та принцип його роботи.

В третьому розділі було розглянуто набір даних для передбачення ядовитості грибів за їх зовнішніми характеристиками. Було розглянуто принципи, що використовувались при розробці моделей, було досліджено вплив гіперпараметрів та порівняно результати якості передбачень з іншими методами.

В четвертому розділі, на основі знань отриманих в розділі 3, було проведено оглядове дослідження ще двох наборів даних, для перевірки та порівняння з результатами основного дослідження.

Джерела

<https://uk.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%80%D0%BE%D0%BC%D1%96%D1%81_%D0%B7%D1%81%D1%83%D0%B2%D1%83_%D1%82%D0%B0_%D0%B4%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D1%96%D1%97>

<https://uk.wikipedia.org/wiki/%D0%A2%D0%BE%D1%87%D0%BD%D1%96%D1%81%D1%82%D1%8C>

<https://uk.wikipedia.org/wiki/F-%D0%BC%D1%96%D1%80%D0%B0>

<https://uk.wikipedia.org/wiki/%D0%93%D1%96%D0%BF%D0%B5%D1%80%D0%BF%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80_(%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%B5_%D0%BD%D0%B0%D0%B2%D1%87%D0%B0%D0%BD%D0%BD%D1%8F)>