**ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ФРАНКА**

Факультет прикладної математики та інформатики

(повне найменування назва факультету)

Кафедра дискретного аналізу та інтелектуальних систем

(повна назва кафедри)

Курсова робота

«Порівняльний аналіз методів класифікації даних»

Виконала: студентка III курсу, групи ПМІ-33 спеціальності

122 Комп'ютерні науки та інформаційні технології

(шифр і назва спеціальності)

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Бень Х. Б. |  |
| Керівник | (підпис) | (прізвище та ініціали)  Квасниця Г. А. |  |
|  | (підпис) | (прізвище та ініціали) |  |

Львів – 2024

Зміст

[Вступ 3](#_Toc136515294)

[Розділ 1. Теоретичні основи класифікації 5](#_Toc136515295)

[1.1 Означення класифікації 5](#_Toc136515296)

[1.2 Види класифікації 7](#_Toc136515297)

[1.3 Процес класифікації 10](#_Toc136515298)

[1.4 Оцінювання класифікації 12](#_Toc136515299)

[1.5 Методи класифікації 16](#_Toc136515300)

[1.6 Означення кластеризації 17](#_Toc136515301)

[Розділ 2. Метод логістичної регресії 19](#_Toc136515302)

[2.1 Теоретичний опис методу 19](#_Toc136515303)

[2.2 Покроковий хід алгоритму 21](#_Toc136515304)

[2.3 Програмна реалізація 22](#_Toc136515305)

[2.4 Аналіз результатів 24](#_Toc136515306)

[Розділ 3. Метод дерева рішень 27](#_Toc136515307)

[2.1 Теоретичні відомості 27](#_Toc136515308)

[2.2 Покроковий хід алгоритму 30](#_Toc136515309)

[2.3 Програмна реалізація 31](#_Toc136515310)

[2.4 Аналіз результатів 32](#_Toc136515311)

[Висновок 36](#_Toc136515312)

[Джерела 38](#_Toc136515313)

Вступ

У наш час наплив інформації завдяки можливостям її зберігання та збирання став, водночас, винятковою цінністю і великою складністю. Велика кількість інформації, можливість її доступного використання відкрили багато нових можливостей та дали поштовх для розвитку в різних сферах. Проте це створило таку проблему, як невміння користуватись даними правильно. Такі обсяги інформації часто видаються складними для інтерпретації. Щоб їх коректно застосовувати потрібно вміти систематизувати для подальшого аналізу. Тому, необхідно вивчати різні способи їхнього впорядкування.

Актуальність теми з практичної точки зору полягає в тому, щоби досліджувати можливості класифікації даних та застосовувати для різних цілей та прикладів.

Зважаючи на вказане, вище обрана тема є актуальною у практичному контексті

Питання класифікації у своїх роботах досліджувало багато вчених. Серед них:

* Френсіс Голтон (Francis Galton) - робота "Класифікація та вбудована асоціація" (Classification and Heredity)
* Рональд Фішер (Ronald Fisher) – праця "Статистичні методи для дослідження різниць та взаємозв'язків" (The Statistical Methods for Research Workers)
* Леонард Айзенберг (Leonard Eisenberg) – "Моделі класифікації та їх використання в розпізнаванні образів" (Classification Models and Their Use in Image Recognition)
* Том Мішель (Tom Mitchell): "Машинне навчання" (Machine Learning)

Оскільки не було знайдено комплексної праці із детальними поясненнями цієї теми, робота є актуальною з теоретичної точки зору.

Таким чином, метою цієї курсової роботи є дослідження загальних класифікаційних понять та особливостей на прикладі двох методів: логістичної регресії та дерева рішень.

Завдання полягає в тому, щоб виконати такі пункти:

* описати що таке класифікація
* визначити її різновиди
* дослідити алгоритми та оцінювання
* охарактеризувати методи
* вивчити метод логістичної регресії
* вивчити метод дерева рішень

Об’єктом дослідження стали методи класифікації, зокрема – логістична регресія та дерево рішень.

Предметом було визначено застосування цих методів для вступної кампанії та рейтингу успішності студентів.

Практичне застосування цих прикладів може бути використано в аналізі наступних вступних кампаній, поділі рейтингових списків або в інших прикладах.

# Розділ 1. Теоретичні основи класифікації

## Означення класифікації

Для глибинного розуміння класифікаційних процесів необхідний детальний розбір принципів та теоретичних понять.

Можна подати декілька визначень для чіткого уявлення узагальненого поняття «класифікації».

Класифікація – це процес розподілення елементів предметної області на конкретні підгрупи за певним алгоритм, що формується визначальними характеристиками.

Класифікація – упорядкована за деяким принципом множина об’єктів, які мають подібні класифікаційні ознаки (одна або декілька властивостей), обраних для визначення схожості або відмінності між цими об’єктами [1].

Класифікація – системний розподіл досліджуваних предметів, явищ, процесів за родами, видами, типами, з якими-небудь істотними ознаками для зручності їхнього дослідження; угрупування вихідних понять і розташування їх у певному порядку, що відбиває ступінь цієї схожості.

Завдання класифікації – це віднесення об’єкту до того чи іншого класу.

Для здійснення класифікації необхідно дотримуватись певних визначених правил:

1. у кожному акті ділення необхідно застосовувати тільки одну основу;
2. ділення повинне бути пропорційним, тобто загальний обсяг видових понять повинен дорівнювати обсягу діленого родового поняття;
3. члени ділення повинні взаємно виключати один одного, їхні обсяги не повинні перетинатися;

4) ділення повинне бути послідовним [2].

## 1.2 Види класифікації

Можна розрізняти такі види класифікації:

1. *за ознаками*
2. **допоміжна (штучна) класифікація** – відбувається з використанням якоїсь зовнішньої ознаки, яка не є суттєвою ознакою, але оптимальна для застосування, щоб надати порядок предметній області. Наприклад: відсортовування учнів за алфавітом – ми можемо швидко відшукати необхідне прізвище, що для деяких завдань є зручно, але це не дає нам інформативного наповнення про вміння та якості особи;
3. **природня класифікація** – це складніший варіант способу впорядкування, оскільки потребує вивчення та аналіз властивостей об’єкту. Завданням є пошук закономірностей між елементами для виокремлення їхніх властивостей. Ґрунтується на природніх зв’язках та подібностях чи характерних особливостях. Також може використовуватися для передбачення, якої категорії будуть стосуватися нові дані. Наприклад: класифікація для формування видів тварин чи рослин у біології [1].
4. *за процедурою розподілу*
5. **проста** – використовує невелику кількість категорій або класів для розподілу об'єктів. Наприклад, може бути використана бінарна класифікація, де об'єкти поділяються на два взаємовиключні класи. Це простий підхід, який дає можливість зрозуміти основний аспект або властивість об'єктів, але може бути недостатнім для більш деталізованого аналізу або розпізнавання складних взаємозв'язків. Може бути ефективною, якщо основна інформація може бути висвітлена в межах обмеженого набору класів;
6. **складна** – використовує більшу кількість категорій або класів для розподілу об'єктів. Цей підхід дозволяє враховувати більшу різноманітність характеристик і відтворювати більш деталізовану структуру або відносини між об'єктами. Складна класифікація може включати багатокласову класифікацію з численними категоріями або природну ієрархічну класифікацію, де класи організовані у вигляді дерева або ієрархії з різними рівнями. Може бути необхідною, коли варто враховувати багато різноманітних характеристик або з'ясувати багатоаспектну залежність між об'єктами.
7. *за кількістю ознак*
8. **одновимірна** – використовує лише одну ознаку або змінну для розподілу об'єктів на класи. Цей підхід простий та прямолінійний, оскільки базується на аналізі лише одного аспекту об'єктів. Наприклад, можна класифікувати студентів на основі їхнього віку, де єдиним критерієм є числове значення віку. Одновимірна класифікація може бути корисною, коли важлива лише одна особливість або коли інші ознаки не мають значення для вирішення конкретного завдання. Тобто є ефективнішою для наборів даних із простим сценарієм;
9. **багатовимірна** – використовує кілька ознак або змінних для розподілу об'єктів на класи. Цей підхід дозволяє враховувати більш широкий набір характеристик об'єктів, що надає більше гнучкості у випадках, коли важливо врахувати багато аспектів. Наприклад, можна класифікувати студентів за віком, середнім балом, кількістю годин, витрачених на вивчення і т. д. Кожна з цих ознак може бути використаною як вимір для розподілу об'єктів у багатовимірному просторі. Багатовимірна класифікація дозволяє отримати більш повну інформацію про зв'язки та структуру даних, а також може покращити точність класифікаційної моделі для оптимальнішого аналізу даних.
10. *за метою:*
11. **прогнозування** – застосовується для передбачення результату від надходження нових даних. Наприклад, прогнозування приросту попиту для певних товарів;
12. **ідентифікація** – вирізнення, якої з категорій стосується конкретний елемент за його характеристиками. Прикладом може бути ідентифікація осіб за їхніми біометричними даними;
13. **систематизація –** використовується для організації та впорядкування даних, дає можливість групувати схожі об'єкти разом, що полегшує подальше оброблення. Наприклад, формування баз даних і створення спрощеної системи пошуку елементів;
14. **виявлення аномалій –** відокремлення елементів від нормального розподілу.Може застосовуватися, щобивідокремити нормальну активність від підозрілої активності в мережі або для виявлення шахрайських операцій у банківських рахунках;
15. **пріоритезація –** необхідна для ранжування об’єктів за їхньою релевантністю. Може використовуватись для формування порядку рекомендацій веб-сторінок після запиту у пошукових системах.

## 3 Процес класифікації

Види наборів вхідних даних вирізняють на дві множини: навчальну і тестову.

**Навчальна множина (training set)** – множина, яка включає дані, що використовуються для навчання (конструювання) моделі. Така множина містить вхідні та вихідні (цільові) значення прикладів. Вихідні значення призначені для навчання моделі.

**Тестова множина (test set)** також містить вхідні та вихідні значення прикладів. Тут вихідні значення використовуються для перевірки працездатності моделі [1].

Варто описати в якому порядку відбувається класифікація. Процес складається з двох етапів: конструювання моделі та її використання.

1. **Конструювання моделі.** На цьому етапі відбуваються підготовчі процеси для формування математичної формули чи тренувальної моделі, до яких пізніше застосовується конкретний алгоритм. Тобто ми описуємо множину визначених класів.

1. ***Збір та підготовка даних***: необхідно зібрати відповідні дані, які будуть використовуватись для класифікації. Це можуть бути числові дані, текстові описи, зображення, аудіо або будь-який інший тип даних. Потім дані підлягають попередньому обробленню, включаючи очищення, перевірку на коректність та масштабування.
2. ***Визначення та пріоритезація ознак***: необхідно вибрати набір ознак, що описують об'єкти, та визначити, які мають найбільше значення. Вибір може відбуватися або вручну, або враховуючи результати обчислення цінності інформації, що несе кожна ознака.
3. ***Вибір алгоритму*:** вибір алгоритму залежить від типу даних та завдання класифікації. Якщо нам необхідно робити вибір, враховуючи велику кількість аспектів, то краще вибирати алгоритми машинного навчання, а якщо ми будемо працювати з математичними обчисленнями – статистичного.
4. ***Тренування моделі***. Після визначення алгоритму необхідно навчити модель на тренувальному наборі даних. Тренування моделі включає пошук оптимальних параметрів алгоритму, які найкраще адаптуються до навчальних даних. Це може включати процес навчання з учителем, де дані мають мітки класів, або навчання без учителя, де дані не мають міток, і модель виявляє внутрішні закономірності або групи в даних.
5. ***Валідація моделі***. Після тренування моделі необхідно перевірити її ефективність та точність на валідаційному наборі даних. Це допомагає оцінити якість класифікації та зробити необхідні налаштування для поліпшення моделі

2. **Використання моделі**. Етап, на якому відбувається власне поділ об’єктів предметної області. Робота відбувається з новими або тестовими даними, до яких ми застосовуємо нашу натреновану модель. Рівень точності – відсоток правильно класифікованих прикладів у тестовій множині.

1. ***Тестування моделі***. Після успішної валідації модель може бути протестована на тестовому наборі даних, які модель раніше не бачила. Це дозволяє оцінити загальну здатність моделі до класифікації нових об'єктів та отримати бажаний результат.
2. ***Оцінка правильності***. Нам необхідно отримати підтвердження коректності результатів. Для цього розраховується рівень точності.
3. ***Застосування моделі***: Після успішного тестування модель може бути застосована для класифікації нових невідомих об'єктів або даних в реальному часі.

Процес класифікації може бути ітеративним, і вимагати налаштування параметрів моделі та оптимізацію для досягнення кращої точності класифікації.

## ***Оцінювання класифікації***

Є чотири типи оцінювання якості моделі класифікації:

1. **Крос-перевірка (Cross-Validation) –** використовує один початковий набір даних, який поділяється на кілька підмножин. Модель навчається на одних підмножинах, а оцінка її ефективності проводиться на інших. Цей процес повторюється кілька разів з різними комбінаціями. В результаті отримується усереднена метрика ефективності моделі. Цей спосіб варто використовувати для вибірок з великою кількістю прикладів так, щоби навчальна і тестова вибірки не могли перетинатися [5].
2. **Тестова множина (Test Set)** – це оцінювання, яка на початку отримує два чітко поділені набори даних: Тестова вибірка – це окремий набір даних, який не використовується під час навчання моделі для обчислення точності. Модель, яка була навчена на тренувальному наборі, використовується для класифікації об'єктів у тестовій множині.
3. **Таблиця спряженості (Confusion Matrix)** – це інструмент для оцінки точності моделі класифікації. Її елементами у головній діагоналі є кількість правильно класифікованих об'єктів для кожного класу та неправильно класифікованих – на побічній.

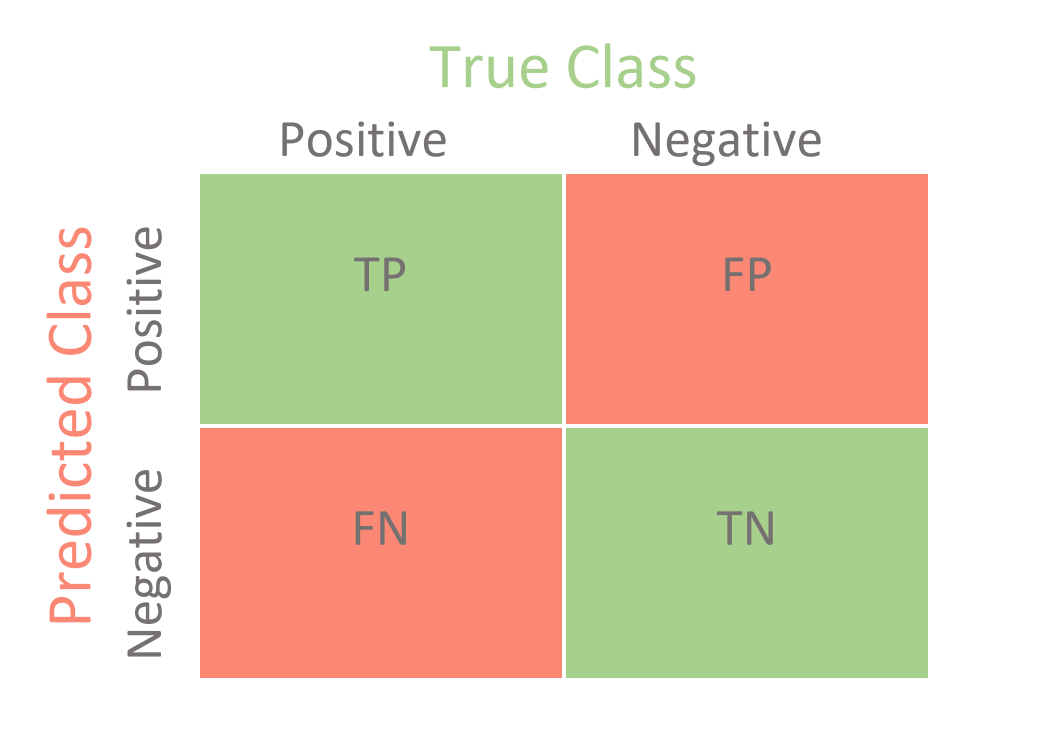


Рис. 1. Таблиця спряженості

* TP (True Positives) – правильно класифіковані зразки (істинно-позитивні випадки);
* TN (True Negatives) — правильно класифіковані негативні приклади (істинно-негативні випадки);
* FN (False Negatives) – позитивні зразки, класифіковані як негативні. Це помилка 1-го роду (хибно-негативні зразки);
* FP (False Positives) – негативні зразки, класифіковані як позитивні. Це помилка 2-го роду (хибно-позитивні випадки) [8].

Як наслідок, з цієї таблиці можна вивести такі показники, як:

* **Точність (accuracy)** – дає загальну точність моделі, тобто частку правильно класифікованих від загальної кількості зразків. Для розрахунку використовують таку формулу:

AC = (TP+TN)/(TP+TN+FP+FN)

* **Коефіцієнт неправильної класифікації (Misclassification Rate)**: повідомляє, яка частка прогнозів була неправильною. Також відомий, як помилка класифікації. Можна обчислити його за допомогою

MR = (FP+FN)/(TP+TN+FP+FN) або (1-AC)

* **Правильність (precision)**: несе в собі інформацію, яка частка прогнозів позитивного класу була насправді позитивною. Щоб обчислити точність, використовують наступну формулу:

PR = TP/(TP+FP)

* **Чутливість (recall/sensitivity)**: дає інформацію про те, яку частку всіх позитивних зразків класифікатор правильно передбачив як позитивні. Він також відомий як **істинний позитивний коефіцієнт (TPR)**. Для обчислення використовують таку формулу:

SE=TPR= TP/(TP+FN)

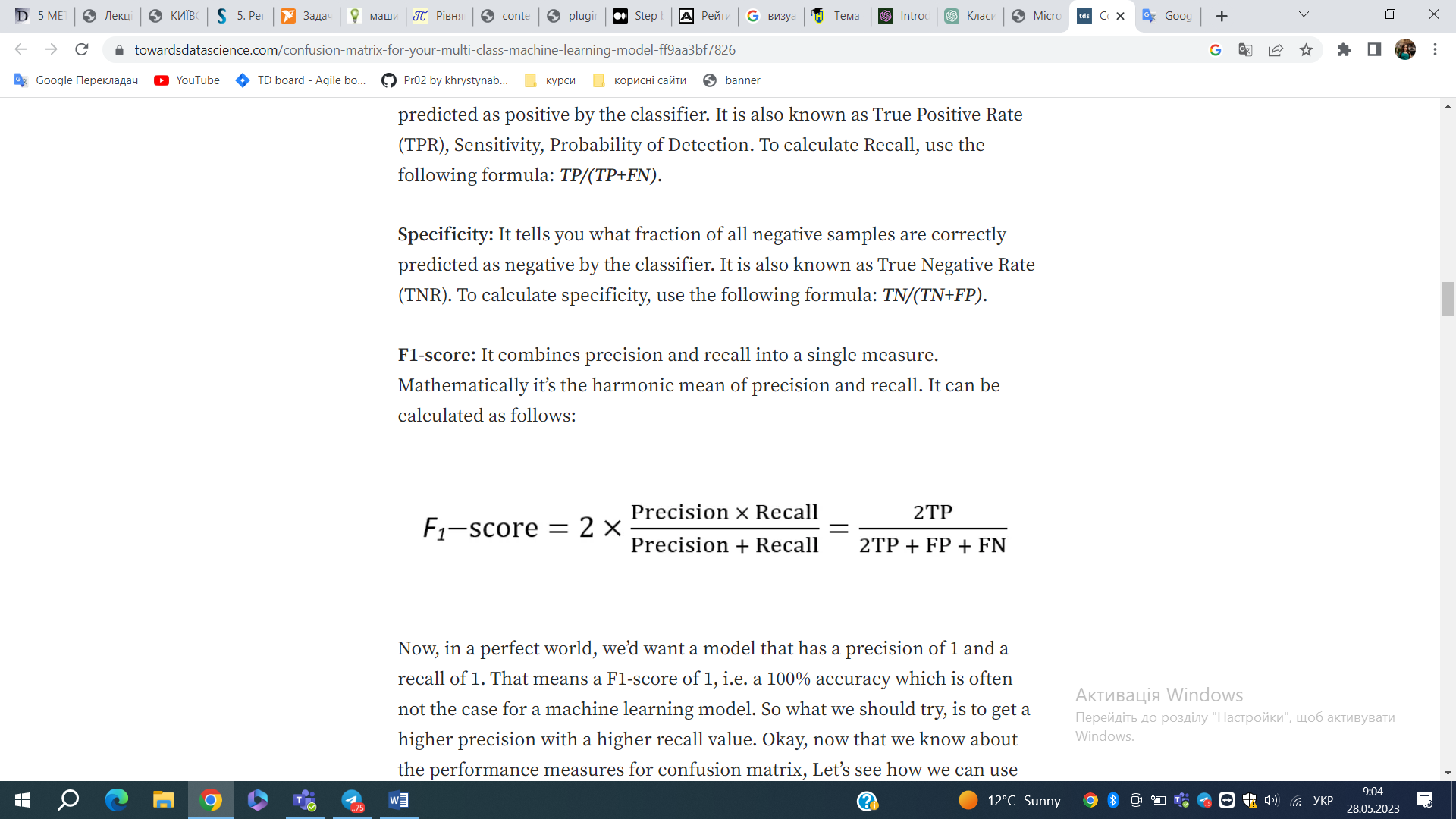
* **Хибний позитивний коефіцієнт (FPR):** частка, яка обчислюється такою формулою:

FPR = FP/(TN+FP)

* **Специфічність (specificity)** - це частка всіх негативних зразків, які класифікатор правильно передбачив як негативні. Він також має назву – справжній **негативний коефіцієнт (TNR).** Для його розрахунку маємо формулу:

SP =TNR=TN/(TN+FP)

* **F1-міра (F1-score)**: поєднує в собі точність і запам’ятовування. Математично – це середнє гармонійне значення точності та запам’ятовування. Його можна розрахувати наступним чином:



1. **ROC-аналіз (Receiver Operating Characteristic)**: відбувається завдяки аналізу результатів внаслідок побудови ROC-кривої. Він показує залежність між чутливістю (доля правильно визначених позитивних об'єктів) та специфічністю (доля правильно визначених негативних об'єктів) за зміни порогу. Крива ROC може бути використана для порівняння різних моделей та вибору оптимального порогу відсічення для заданої задачі класифікації.

ROC-крива будується як графік, де на осі X відображається специфічність, а на осі Y – чутливість. Для кожного порогового значення обчислені метрики відображаються на графіку, і відрізок з'єднує ці точки.

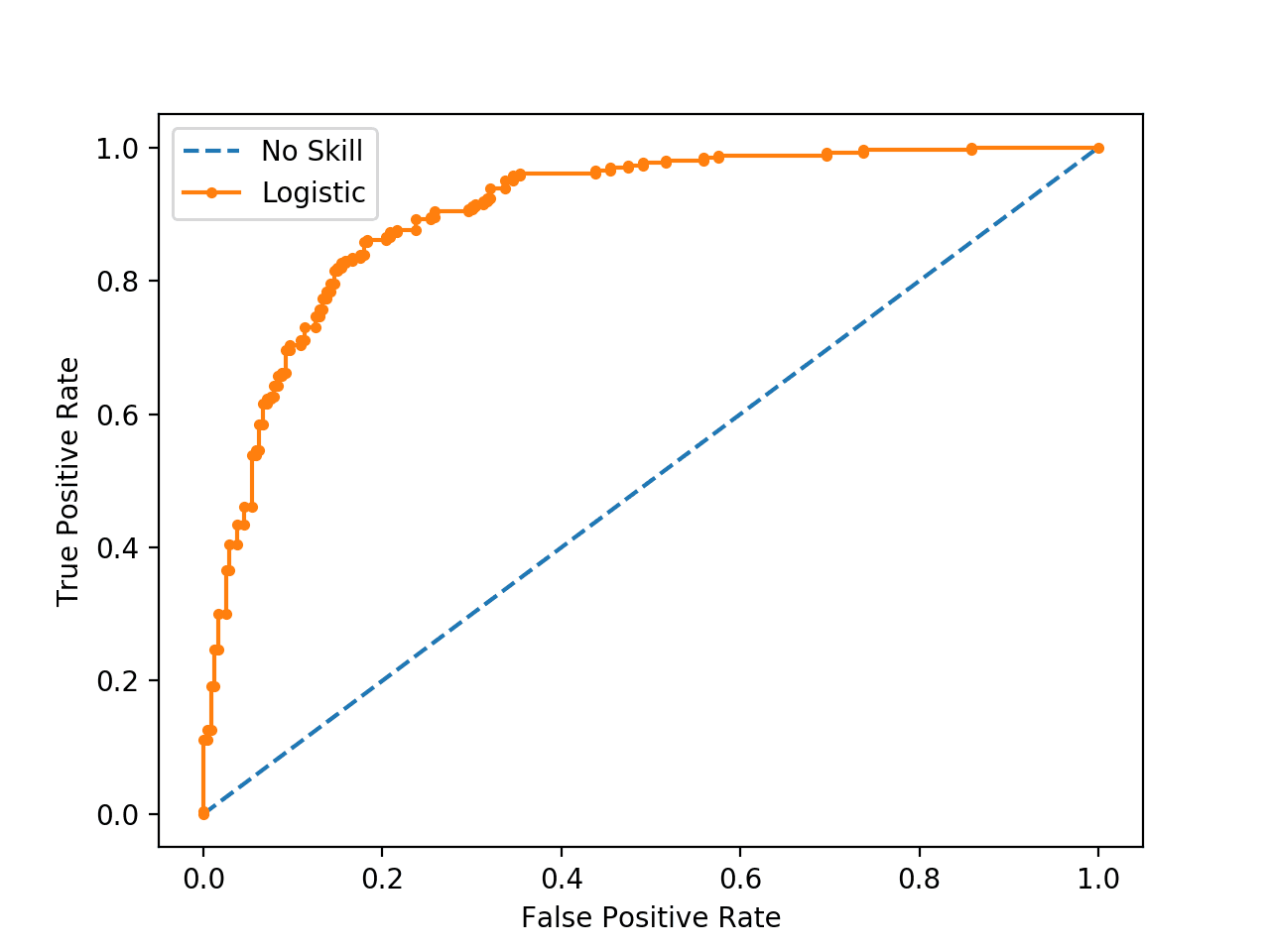
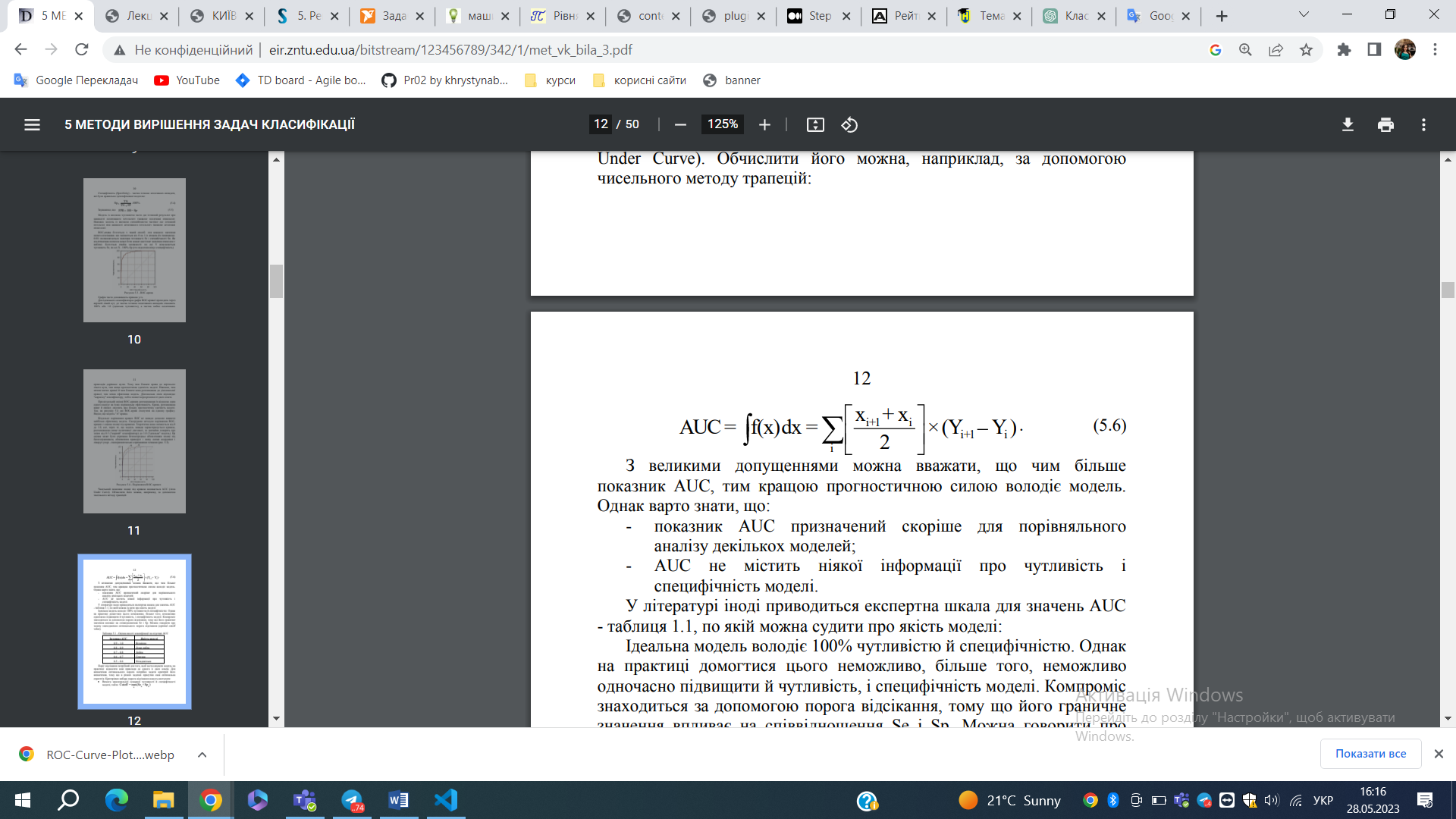


Рис. 2. ROC-крива

Точність може бути оцінена за допомогою площі під кривою (Area Under the Curve - AUC). Чим більше AUC, тим краще модель може розрізнити класи [7].



## Методи класифікації

Методи умовно поділяють на дві групи.

1. Статистичні методи класифікації

* байєсівська класифікація;
* логістична регресія;
* дискримінантний аналіз

1. Методи машинного навчання для класифікації

* за допомогою дерев рішень;
* класифікація за допомогою штучних нейронних мереж;
* класифікація за допомогою алгоритмів покриття;
* класифікація методом опорних векторів;
* класифікація за допомогою методу k-найближчих сусідів;
* класифікація CBR-методом [2]

Оцінювання методів слід проводити, виходячи з таких характеристик:

1. **Швидкість** характеризує час, який потрібен на створення моделі та її використання.
2. **Робастність**, тобто стійкість до будь-яких порушень вихідних передумов, означає можливість роботи з зашумленими даними і пропущеними значеннями в даних.
3. **Інтерпретованість** забезпечує можливість розуміння моделі аналітиком.
4. **Надійність** методів класифікації передбачає можливість роботи цих методів при наявності в наборі даних шумів і викидів.

## 1.6 Означення кластеризації

Такий тип сортування об’єктів, що має наперед визначені класи та повинен розрізнити елементи між ними, називають «класифікація з учителем».

Проте не завжди ми отримуємо сформовані класи, за якими сортувати.

В такому випадку ми отримуємо задачу «класифікації без учителя». Відповідно нам потрібно сформувати класи, розділивши заданий набір об’єктів на групи за певними критеріями, що означає «здійснити кластерний аналіз».

Кластеризація – це метод аналізу даних, який використовується для групування схожих об'єктів або прикладів у наборі даних.

Основна ідея полягає в тому, щоб об'єкти, які схожі один на одного, були розташовані в одному кластері, тоді як об'єкти, які відрізняються, були розташовані у різних кластерах.

Метою кластеризації є пошук існуючих структур.

Кластер можна охарактеризувати як групу об’єктів, що мають загальні властивості. Характеристиками кластера можна назвати дві ознаки:

* внутрішня однорідність;
* зовнішня ізольованість.

Наведемо коротку характеристику підходів до кластеризації.

1. Алгоритми, засновані на поділі даних (Partitioning algorithms), у тому числі ітеративні:

• поділ об’єктів на k кластерів;

• ітеративний перерозподіл об’єктів для поліпшення кластеризації.

1. Ієрархічні алгоритми (Hierarchy Algorithms):

• агломерація: кожен об’єкт спочатку є кластером, кластери, об’єднуючись один з одним, формують більший кластер і т. д.

1. Методи, засновані на концентрації об’єктів (Density-based methods):

• засновані на можливості з’єднання об’єктів;

• ігнорують шуми, для знаходження кластерів довільної форми.

1. Грід-методи (Grid-based methods):

• квантування об’єктів в грід-структури.

1. Модельні методи (Model-based):

• використання моделі для знаходження кластерів, найбільш відповідних даним [1]

Оцінка якості кластеризації може бути проведена на основі таких процедур:

• ручна перевірка;

• встановлення контрольних точок та перевірка на отриманих кластерах;

• визначення стабільності кластеризації шляхом додавання в модель нових змінних;

• створення та порівняння кластерів із використанням різних методів. Різні методи кластеризації можуть створювати різні кластери, і це є нормальним явищем. Однак створення схожих кластерів різними методами вказує на правильність кластеризації.

Застосування кластерного аналізу:

* у медицині використовується кластеризація захворювань, лікування захворювань або їх симптомів;
* в археології встановлюються таксономії кам’яних споруд і стародавніх об’єктів і т. д.;
* у маркетингу це може бути задача сегментації конкурентів і споживачів;
* у менеджменті прикладом задачі кластеризації буде розбиття персоналу на різні групи, класифікація споживачів і постачальників, виявлення схожих виробничих ситуацій, під час яких виникає брак;
* у соціології задача кластеризації – розбиття респондентів на однорідні групи.

# Розділ 2. Метод дерева рішень

## Теоретичні відомості

Метод дерев рішень (decision trees) є одним із найбільш популярних методів розв’язання задач класифікації й прогнозування. Іноді цей метод Data Mining також називають деревами вирішальних правил, деревами класифікації і регресії.

За допомогою даного методу розв’язуються задачі класифікації й прогнозування.

Якщо залежна, тобто цільова змінна приймає дискретні значення, за допомогою методу дерева рішень розв’язується задача класифікації.

Якщо ж залежна змінна приймає безперервні значення, то дерево рішень установлює залежність цієї змінної від незалежних змінних, тобто розв’язує задачу чисельного прогнозування.

У найбільш простому вигляді дерево рішень – це спосіб показу правил в ієрархічній, послідовній структурі. Основа такої структури – відповіді «Так» або «Ні» на низку питань [9].

Бінарні дерева є найпростішим, частковим випадком дерев рішень. У вузлах бінарних дерев розгалуження може відбуватися тільки у двох напрямках, тобто існує можливість тільки двох відповідей на поставлене питання («так» і «ні»). В інших випадках, відповідей і, відповідно, гілок дерева, що виходять із його внутрішнього вузла, може бути більше двох.

На етапі побудови моделі, власне, і будується дерево класифікації або створюється набір якихось правил.

На етапі використання моделі побудоване дерево, або шлях від його кореня до однієї з вершин, що є набором правил для конкретного клієнта, використовується для відповіді на поставлене питання «Чи видавати кредит?»

Правилом є логічна конструкція, представлена у вигляді «якщо : то :».

Кожна гілка дерева, що йде від внутрішнього вузла, відзначена предикатом розщеплення, який може відноситися лише до одного атрибуту розщеплення даного вузла. Характерна риса предикатів розщеплення: кожний запис використовує унікальний шлях від кореня дерева тільки до одного вузла-розв’язку. Об’єднана інформація про атрибути розщеплення й предикати розщеплення у вузлі називається критерієм розщеплення (splitting criterion).

*Переваги дерева рішень:*

1. Легко зрозуміти і інтерпретувати: Дерева рішень надають простий і зрозумілий спосіб прийняття рішень. Вони можуть бути інтерпретовані і пояснені людям без спеціальних знань в галузі машинного навчання.
2. Використовується для класифікації та регресії: Дерева рішень можуть використовуватись як для задач класифікації, так і для задач регресії. Вони можуть прогнозувати категоріальні мітки або числові значення.
3. Може працювати зі змішаними типами даних: Дерева рішень можуть обробляти набори даних, які містять як категоріальні, так і числові ознаки. Вони можуть автоматично обробляти такі типи даних без потреби в додаткових попередніх перетвореннях.

*Недоліки дерева рішень:*

1. Схильність до перенавчання: Дерева рішень можуть стати складними і перенавчатися на тренувальних даних, особливо якщо глибина дерева занадто велика або використовуються інші параметри, які сприяють складності моделі. Це може призводити до поганої універсальності і низької здатності до узагальнення на нові дані.
2. Вразливість до шуму та змін в даних: Дерева рішень можуть бути чутливими до шуму та випадкових змін в даних. Навіть невеликі зміни в тренувальних даних можуть призводити до істотних змін у структурі та правилах прийняття рішень, що може негативно впливати на результат.

Для розв’язування дерева рішення використовується багато алгоритмів, таких як:

* ID3,
* CART,
* C4.5,
* C5.0,
* NewId,
* ITrule,
* CHAID,
* CN2

У роботі було використано ID3, тому пояснення нього наведені нижче:

**ID3 (Iterative Dichotomizer 3)** – алгоритм дозволяє працювати тільки з дискретною цільовою змінною, тому дерева рішень, побудовані за допомогою цього алгоритму, є такими, що класифікують. Число нащадків у вузлі дерева не обмежене. Не може працювати з пропущеними даними. У основі цього алгоритму лежить поняття інформаційної ентропії - тобто, міра невизначеності інформації. Для того, щоб визначити наступний атрибут, необхідно підрахувати ентропію усіх невикористаних ознак відносно тестових зразків і вибрати той, для якого ентропія мінімальна. Цей атрибут і вважатиметься найбільш доцільною ознакою класифікації [10].

## Покроковий хід алгоритму

*Опис процесу побудови дерева:*

1. Передача початкових даних - набору ознак (X) і відповідних міток або класів (y).
2. Обчислення ентропії (міри неоднорідності) вихідного набору даних.
3. Якщо умова зупинки виконана (наприклад, досягнуто максимальної глибини дерева або мінімальної кількості зразків у листках), створення листка з найчастішим значенням міток.
4. Інакше, для кожної ознаки:

* Обчислення інформаційного приросту, що вимірює, наскільки добре ознака розбиває набір даних.
* Вибір ознаки з найбільшим інформаційним приростом.
* Розбиття набору даних на дві підмножини залежно від значення обраної ознаки.
* Рекурсивний виклик алгоритму для кожної підмножини з новими даними та позначками.

1. Побудова дерева шляхом створення вузлів та листків залежно від результатів кроків 3 і 4.
2. Повернення побудованого дерева.

*Опис процесу передбачення*:

1. Приймання набору ознак нового зразка, для якого потрібно зробити передбачення.
2. Початок з кореня дерева і перевірка значення ознаки, яка відповідає кореневому вузлу.
3. Перехід до наступного вузла, який відповідає значенню ознаки нового зразка.
4. Повторення кроків 2-3 до досягнення листка (кінцевого вузла) дерева.
5. Повернення значення, яке знаходиться в листку, яке представляє передбачену мітку або клас для нового зразка.

## Програмна реалізація

Для реалізації програми було використано три бібліотеки:

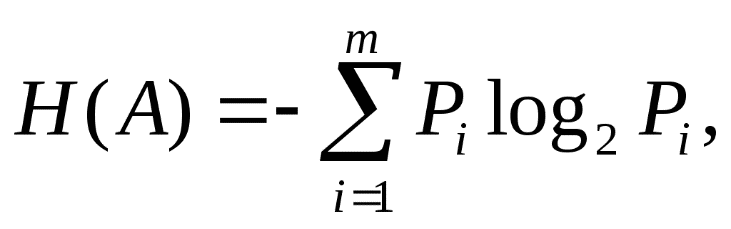
* *numpy* – для операцій з масивами
* *matplotlib* – для побудови графіку
* *pandas* - для роботи з даними
* *pydot* – для візуалізації дерева

1. на початку було підготовлено дані для подальшої роботи:

* *train\_data* – для тренувальної вибірки
* *test\_data* – для тестової вибірки

1. наступним етапом було написання функцій

* *calc\_total\_entropy() –* обчислення початкового значення ентропії для подальших рекурсивних обчислень



(у формулі використано такі позначення: P – ймовірність, m – кількість елементів)

* *calc\_entropy() –* функція, що обчислює ентропію кожного класу

1. знаходження необхідних даних у функціях

* *calc\_info\_gain() –* для обчислення кількості інформації. Це число буде впливати на вибір характеристики, що матиме найбільшу цінність

*–* кількість інформації

*– цінність інформації*

* *find\_most\_inf\_feature() –* відбувається процес пошуку найціннішої характеристики для побудови гілок

1. також було запрограмовано

* *generate\_sub\_tree() –* функція, в якій вузли додаються до дерева
* *make\_tree() –* процес створення дерева зі всіма вузлами, включаючи і гілки(елементи характеристик), і листки (результуючі елементи)

1. *predict() –* проходження по дереву вже з даними тренувальної вибірки
2. створення функції *evaluate()* – для обчислення точності отриманих результатів
3. *plot\_decision\_tree() –* для візуалізації дерева у форматі новоствореного зображення

## 2.4 Аналіз результатів

Для демонстрації роботи алгоритму було використано ті самі дві предметні області з метою отримання таких результатів:

**Предметна область №1:**

1. *Побудова дерева для віднесення студентів до різних груп успішності*

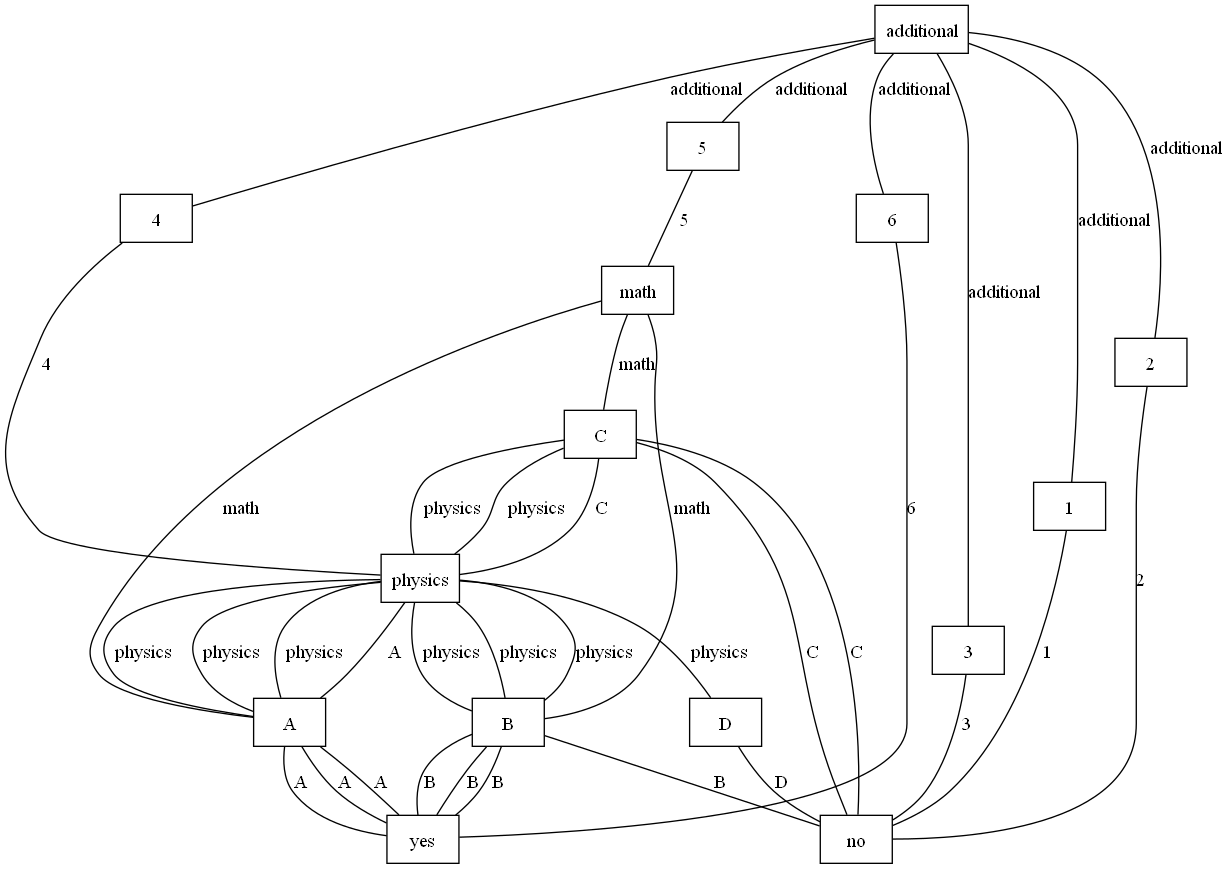
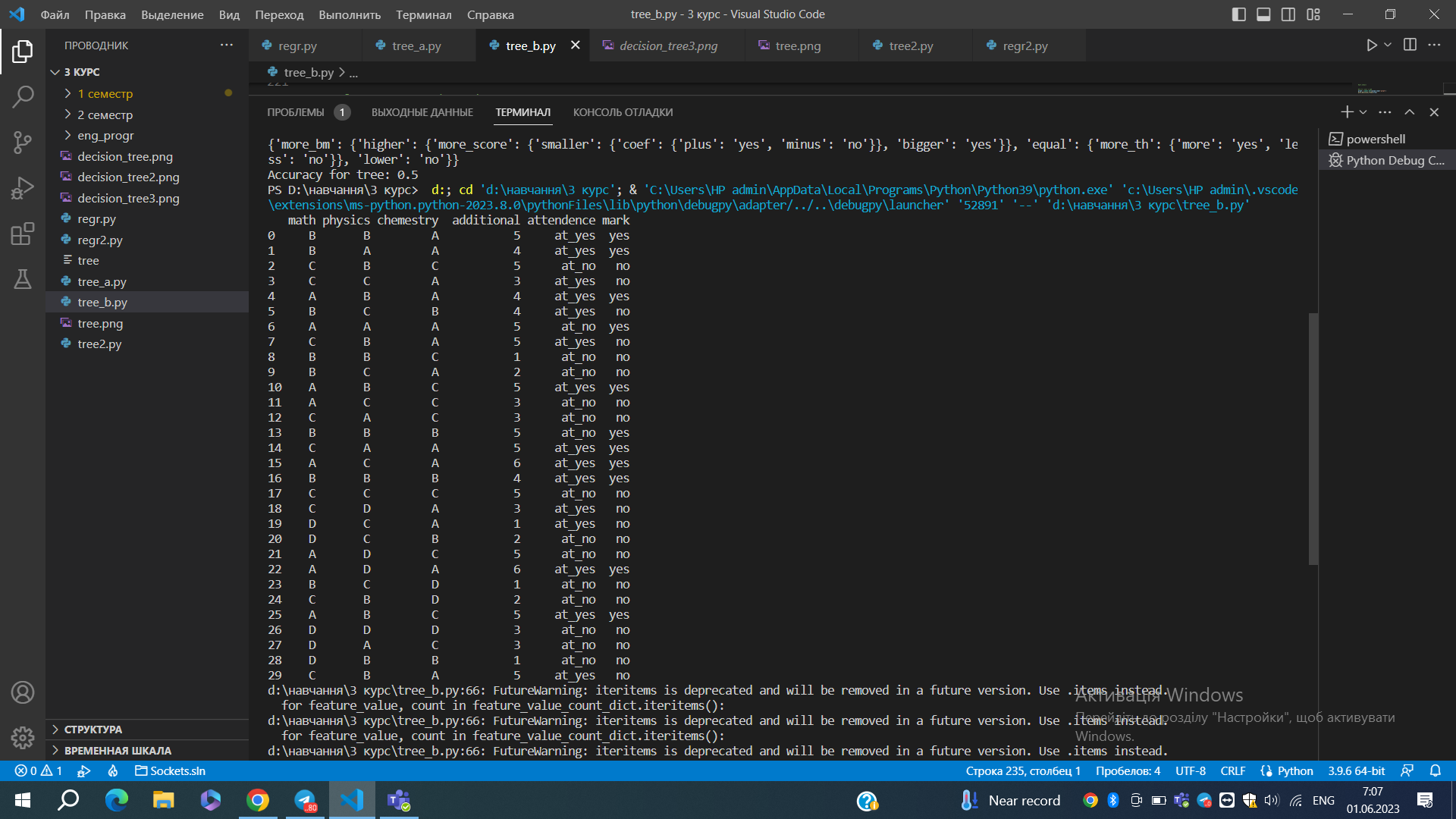


Рис 6. Дерево рішень для предметної області 1

Дерево було побудованими за такими даними:



Де характеристики мають такі значення:

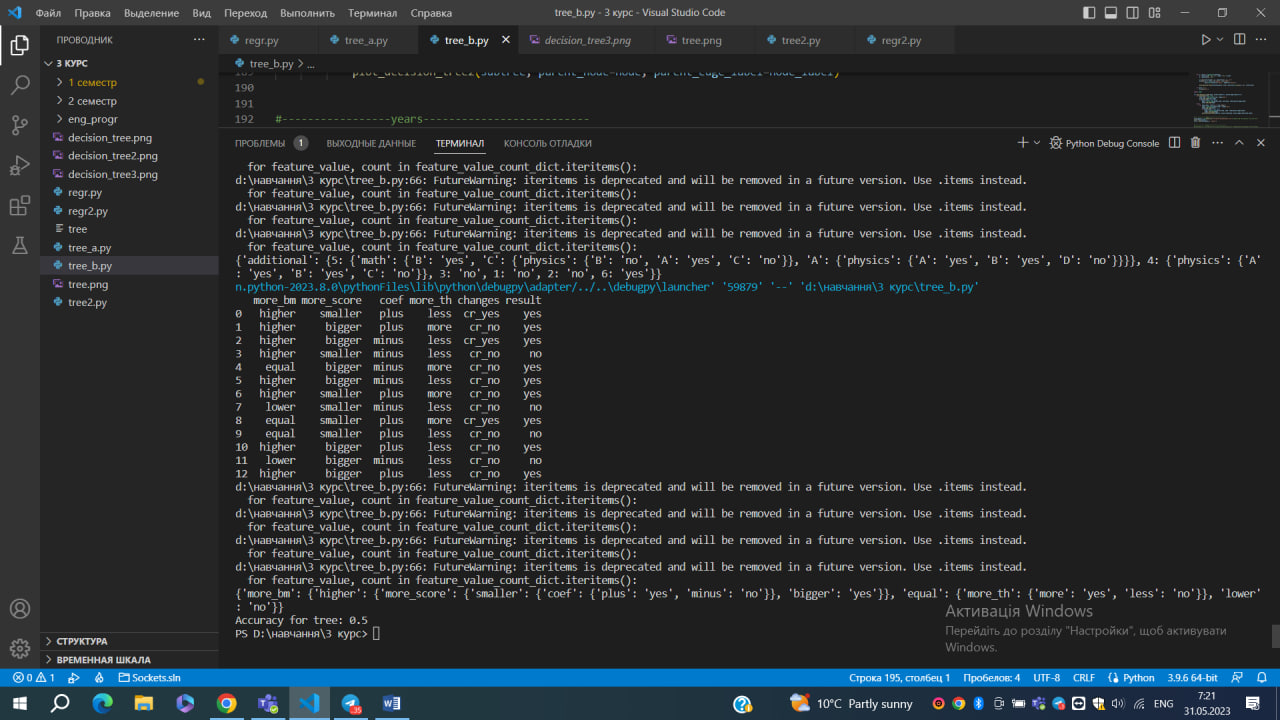
* Оцінка за предмет 1
* Оцінка за предмет 2
* Оцінка за предмет 3
* Наявність додаткових балів
* Пройдений відсоток відвідуваності

Дерево – як отриманий результат, дало можливість завдяки покроковим розгалуженням дізнаватись, куди будуть відноситися нові дані.

Даний приклад демонструє, що наявність додаткових балів мають велику цінність для рейтингу. Проте не має однозначної відповідності, оскільки наступним етапом для визначення є бали за три предмети. Відповідно останнім і найменш суттєвим елементом вважається відвідуваність студентів.

**Предметна область №2:**

1. *Побудова дерева для поділу років вступної кампанії на більш та менш успішні для таких даних*



Де характеристики мають такі значення:

* Bm\_yes – чи збільшилась кількість бюджетних місць
* More\_score – чи збільшився прохідний бал
* Coef – чи більше/менше число, що означає кількість людей на одне бюджетне місце
* More\_th – чи більша кількість заяв
* Changes – чи були зовнішні ускладнюючі фактори
* Result - результат

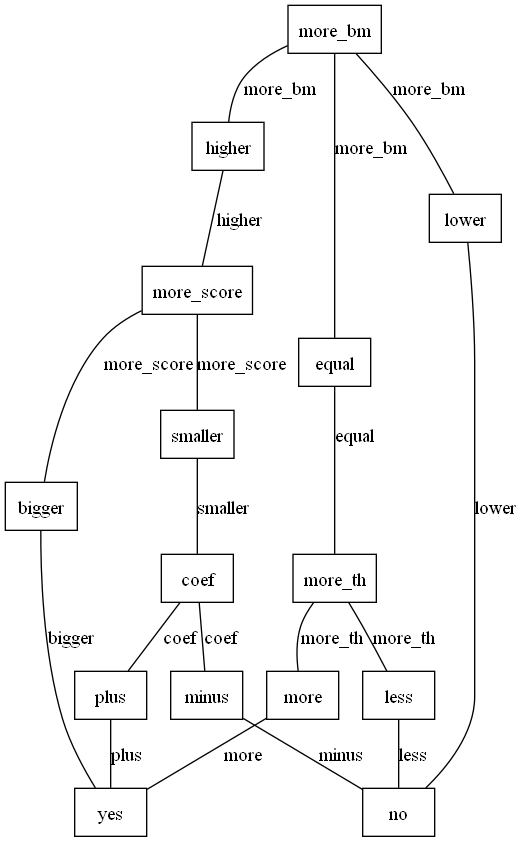


Рис 7. Дерево рішень для предметної області 2

В результаті роботи над тренувальною вибіркою було отримано дерево, що вибудовує схему прийняття рішення, щодо вдалого набору. Отже, найбільш цінною інформацією було визначено характеристику про збільшення бюджетних місць. Наступне – це збільшення прохідного балу у порівнянні із попереднім роком. Приблизно рівноцінними вважаються елементи про кількість та коефіцієнт поданих заяв. А найменш суттєвим показником визначено зміну зовнішніх обставин.

Висновок

У цій роботі було проведено дослідження та порівняння методів класифікації: логістичної регресії і дерев рішень.

Логістична регресія є статистичним методом, який використовує логістичну функцію для прогнозування ймовірності належності об'єкта до певного класу. Вона є простою та ефективною моделлю, яка може добре працювати з числовими та категоріальними ознаками. Логістична регресія також здатна враховувати взаємодію між ознаками та використовувати регуляризацію для уникнення перенавчання. Однак, вона може бути обмежена у відображенні складних залежностей між ознаками та вирішенні нелінійних задач класифікації.

Дерева рішень, з іншого боку, є гнучкими моделями, які використовують набір правил для прийняття рішень. Вони можуть обробляти як числові, так і категоріальні ознаки і мають високу інтерпретованість. Дерева рішень добре підходять для вирішення задач класифікації зі складною структурою даних і можуть бути ефективними для виявлення важливих ознак. Однак, вони можуть схильні до перенавчання, особливо при побудові глибоких дерев, і вони можуть бути вразливі до шуму та змін в даних.

Залежно від характеру задачі та особливостей даних, логістична регресія або дерева рішень можуть бути більш підходящими методами класифікації.

Якщо нам потрібно провести кластеризацію, то необхідно використовувати логістичну регресію, яка також зручна для прогнозування та динаміку розвитку даних з певної предметної області.

Якщо завданням є прогнозування для класифікації нових даних, то дерево рішень буде більш оптимальним, оскільки приймає різні характеристики за типами даних та за суттю. Також варто зазначити, що дерево рішень має в порівнянні більшу точність у прогнозуванні.

Нижче наведено результати дослідження точності двох методів:

|  |  |  |
| --- | --- | --- |
|  | Логістична регресія | Дерево рішень |
| Предметна область про успішність | 0.5 | 1.0 |
| Предметна область про вступну кампанію | 0.67 | 1.0 |

Варто враховувати обмеження та переваги кожного методу при виборі найкращого підходу для конкретного завдання. Крім того, комбінування різних методів може привести до покращення результатів класифікації шляхом використання їх комбінованих переваг.

# Джерела

1. Задачі Data Mining. Класифікація та кластеризація [Електронний ресурс] Режим доступу до ресурсу: https://moodle.znu.edu.ua/pluginfile.php /486129/mod\_resource/content/1/%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D1%8F%206.pdf
2. Інформаційні системи та технології в управлінні. [Електронний ресурс] Режим доступу до ресурсу: http://eir.zntu.edu.ua/bitstream/123456789/342/ 1/met\_vk\_bila\_3.pdf
3. Класифікація алгоритмів машинного навчання: лінійна регресія, класифікація та кластеризація [Електронний ресурс] Режим доступу до ресурсу: https://bloginnovazione.it/uk/machine-learning/3716/
4. Рівняня регресії У на Х та Х на У. Алгоритм та побудова [Електронний ресурс] Режим доступу до ресурсу: https://yukhym.com/uk/vipadkovi-velichini/rivniania-rehresii-alhorytm-ta-pobudova.html
5. Логістична регресія[Електронний ресурс] Режим доступу до ресурсу: http://om.univ.kiev.ua/users\_upload/15/upload/file/lecture\_sm\_08.pdf
6. Огляд логістичної регресії [Електронний ресурс] Режим доступу до ресурсу: https://zephyrnet.com/uk/%D0%BE%D0%B3%D0%BB%D1%8F%D0%B4-%D0%BB%D0%BE%D0%B3%D1%96%D1%81%D1%82%D0%B8%D1%87%D0%BD%D0%BE%D1%97-%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%96%D1%97/
7. How to Use ROC Curves and Precision-Recall Curves for Classification in Python [Електронний ресурс] Режим доступу до ресурсу: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/
8. Confusion Matrix for Your Multi-Class Machine Learning Model [Електронний ресурс] Режим доступу до ресурсу: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826
9. Методи дерев рішень, класифікації та прогнозування [Електронний ресурс] Режим доступу до ресурсу: https://moodle.znu.edu.ua/pluginfile.php?file=/486136 /mod\_resource/content/1/%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D1%8F%209.pdf
10. Дерева рішень і алгоритми їх побудови [Електронний ресурс] Режим доступу до ресурсу: https://ekmair.ukma.edu.ua/server/api/core/bitstreams/590b3734-7e5c-48cc-80ec-f1704589696b/content