Дерева рішень і алгоритми їх побудови

Наквасюк Василь

01

Постановка задачі

Задачі класифікації є досить поширеною в машинному навчанні і мають багато місць використання. Дерева рішень, як один із класифікаторів, успішно використовуються на практиці в таких областях, як:

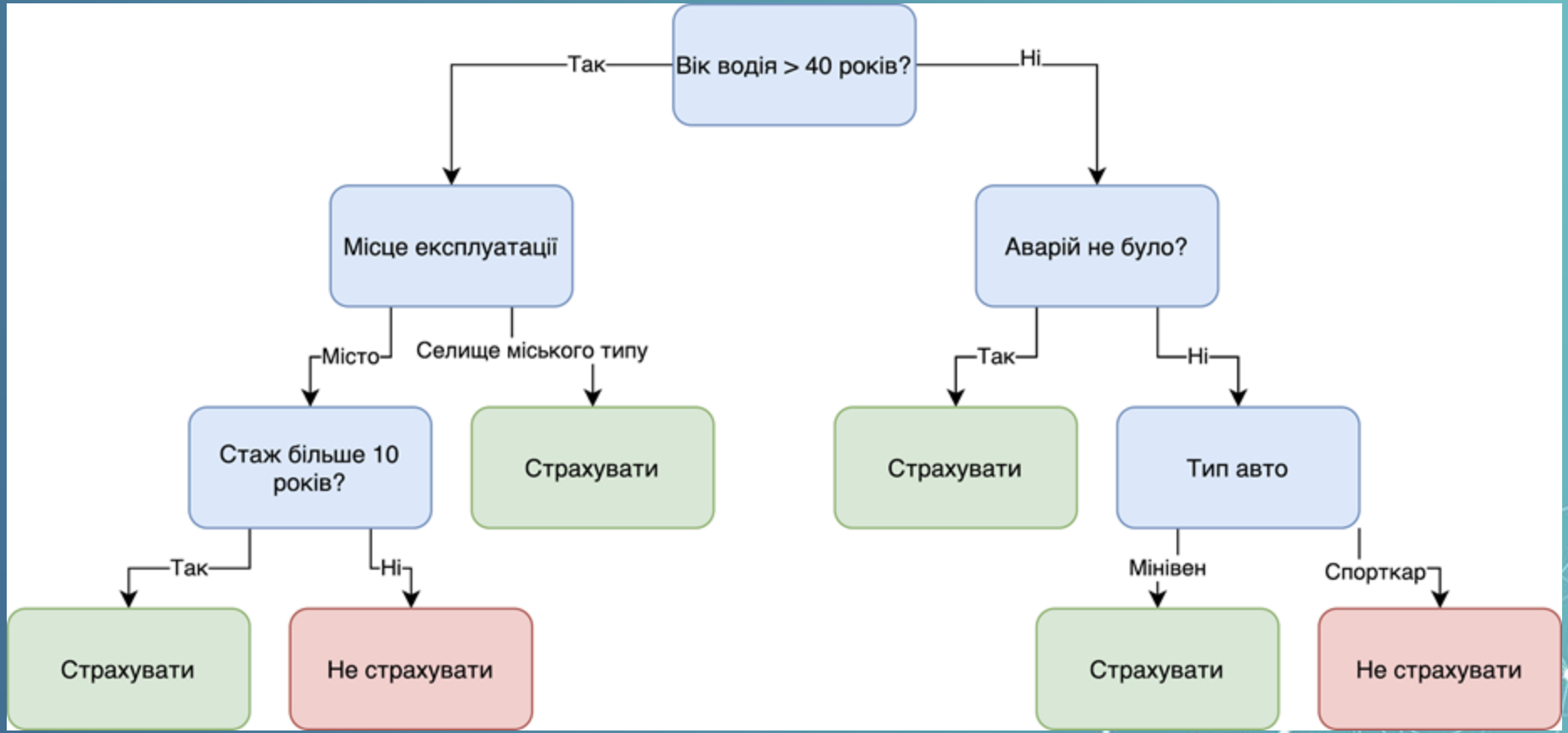
- Банківська справа (оцінка кредитоспроможності клієнтів банку при видачі кредитів)
- Промисловість (контроль за якістю продукції (виявлення дефектів), випробування без руйнувань (наприклад, перевірка якості зварювання) і т.д.)
- Медицина (діагностика захворювань)
- Молекулярна біологія (аналіз будови амінокислот)
- Торгівля (класифікація клієнтів і товарів)

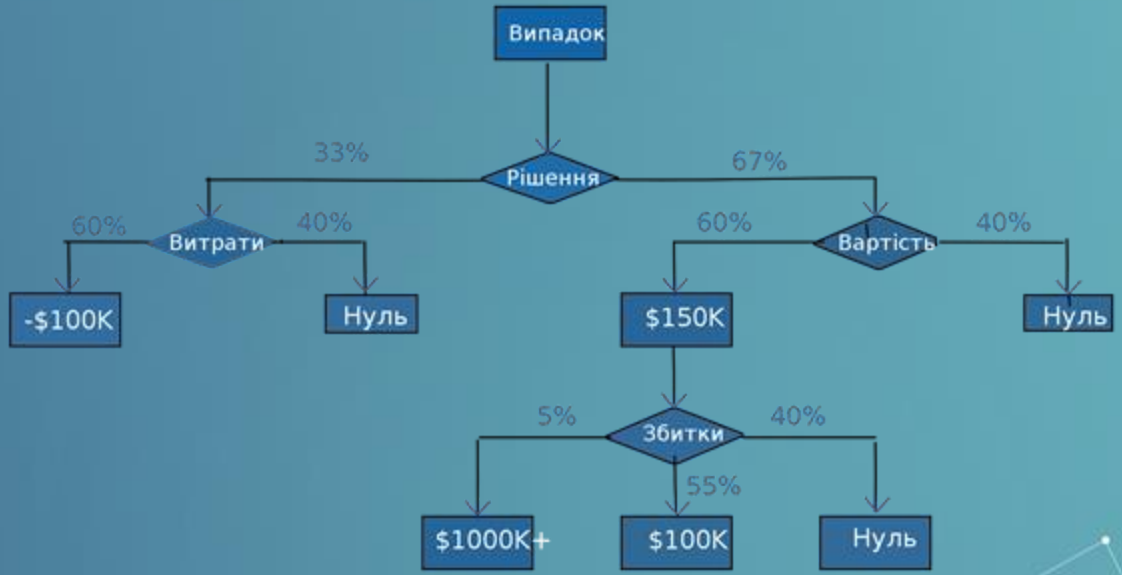
02

Що таке дерево рішень?

- Дерева рішень є одним з найбільш ефективних інструментів інтелектуального аналізу даних і аналітики, які дозволяють вирішувати завдання класифікації і регресії.

- Вони є ієрархічними деревовидними структурами, що складаються з правил рішень виду "Якщо ..., тоді .". Правила автоматично генеруються в процесі навчання на навчальній множині і, оскільки вони формулюються практично на природній мові (наприклад, "Якщо об'єм продажів більше 1000 шт., то товар перспективний"), дерева рішень як аналітичні моделі більш зрозумілі і прозорі, чим, скажімо, нейронні мережі.


Вік водія > 40 років?
Так
Ні
Місце експлуатації
Аварій не було?
Місто
Селище міського типу
Так
Ні
Стаж більше 10 років?
Страхувати
Страхувати
Тип авто
Так
Ні
Мінівен
Спорткар
Страхувати
Не страхувати
Страхувати
Не страхувати



Передбачуваний дохід
Передбачаючи, що вартість, щоб продовжити $50K Передбачаючи, що рішення коштує $100K з комерційних причин =67% x $1COK+67% x ($150k+(5% x $1000K+55% x $100K))-33% x 60% x $100K - $50K - 33% x $100K =67K

ЗАСТОСУВАННЯ

- Класифікація - віднесення об'єктів до одного із заздалегідь відомих класів. Цільова змінна повинна мати дискретні значення.

- Регресія — визначення числового значення незалежній змінній для заданого вхідного вектору.

- Опис об'єктів — набір правил в дереві рішень дозволяє компактно описувати об'єкти. Тому замість складних структур, що описують об'єкти, можна зберігати дерева рішень.

АЛГОРИТМИ ПОБУДОВИ

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.

C4.5 – вдосконалена версія алгоритму ID3, в яку додана можливість роботи з пропущеними значеннями атрибутів.

CART (Classification and Regression Tree) – алгоритм навчання дерев рішень, що дозволяє використати як дискретну, так і безперервну цільову змінну, тобто вирішувати як завдання класифікації, так і регресії.

03

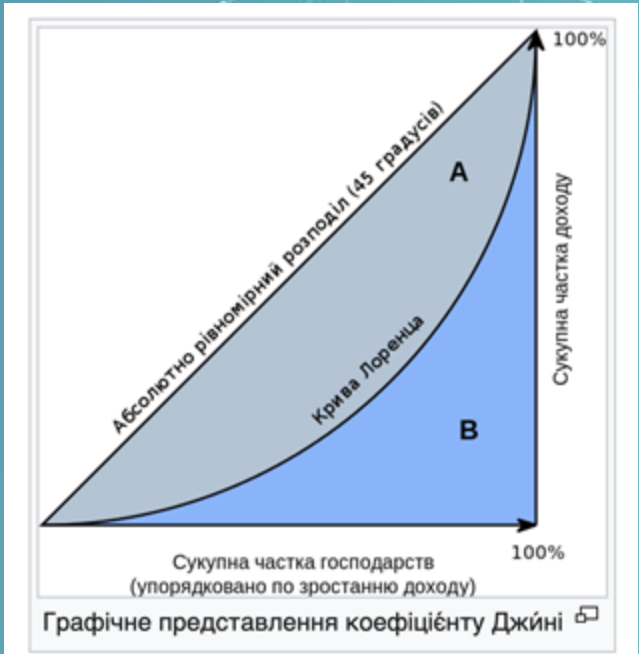
Реальний приклад застосування

ЧИ БУДЕ КЛІЄНТ ПЛАТНИМ?

Звідки прийшов?	Країна	Читав FAQ?	Скільки проглянув сторінок?	Який пакет послуг вибрав?
Slashdot	США	Так	18	Ні
Google	Франція	Так	23	Преміум
Digg	США	Так	24	Базовий
Kiwitobes	Франція	Так	23	Базовий
Google	Великобританія	Ні	21	Преміум
direct	Нова Зеландія	Ні	12	Ні
direct	Великобританія	Ні	21	Базовий
Google	США	Ні	24	Преміум
Slashdot	Франція	Так	19	Ні
Digg	США	Ні	18	Ні
Google	Великобританія	Ні	18	Ні
…	…	…	…	…

КОЕФІЦІЄНТ ДЖИНІ

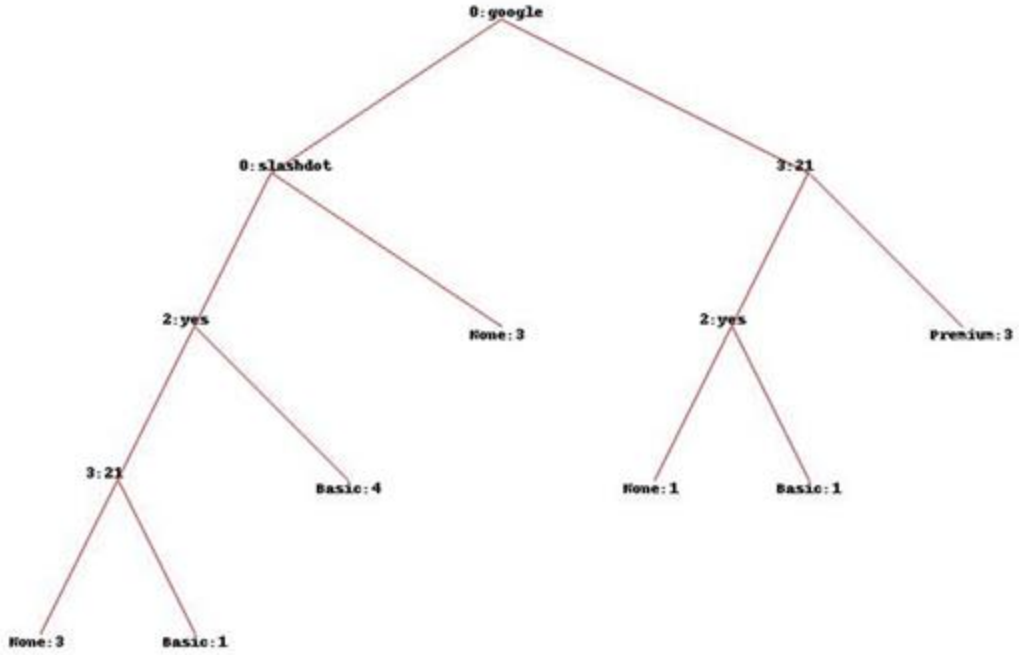
Коефіцієнт Джині — показник нерівності розподілу деякої величини, що приймає значення між 0 і 1, де 0 означає абсолютну рівність (величина приймає лише одне значення), а 1 позначає повну нерівність.



Графічне представлення коефіцієнту Джині

РЕЗУЛЬТАТ

Тоді наше дерево буде мати вигляд:



ВИКОРИСТАННЯ КЛАСИФІКАТОРА

```
def classify(observation,tree):
  if tree.results!=None:
    return tree.results
  else:
    v=observation[tree.col]
    branch=None
    if isinstance(v,int) or isinstance(v,float):
      if v>=tree.value: branch=tree.tb
      else: branch=tree.fb
    else:
      if v==tree.value: branch=tree.tb
      else: branch=tree.fb
    return classify(observation,branch)
```

```
>>> treepredict.classify(['(direct)','USA','yes',5],tree)
{'Basic': 4}
```

Дякую за увагу

Ваші запитання?

v.nakvasiuk@ukma.edu.ua