

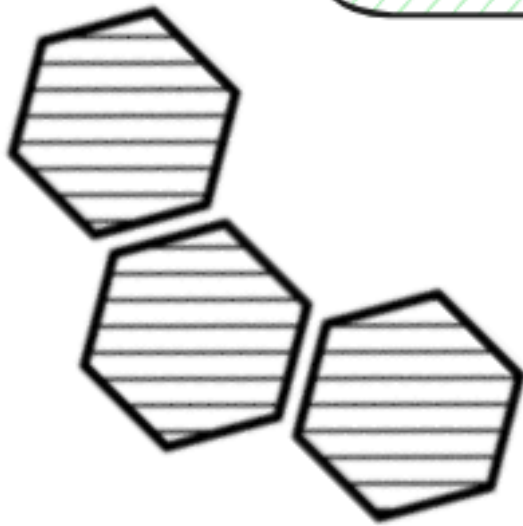
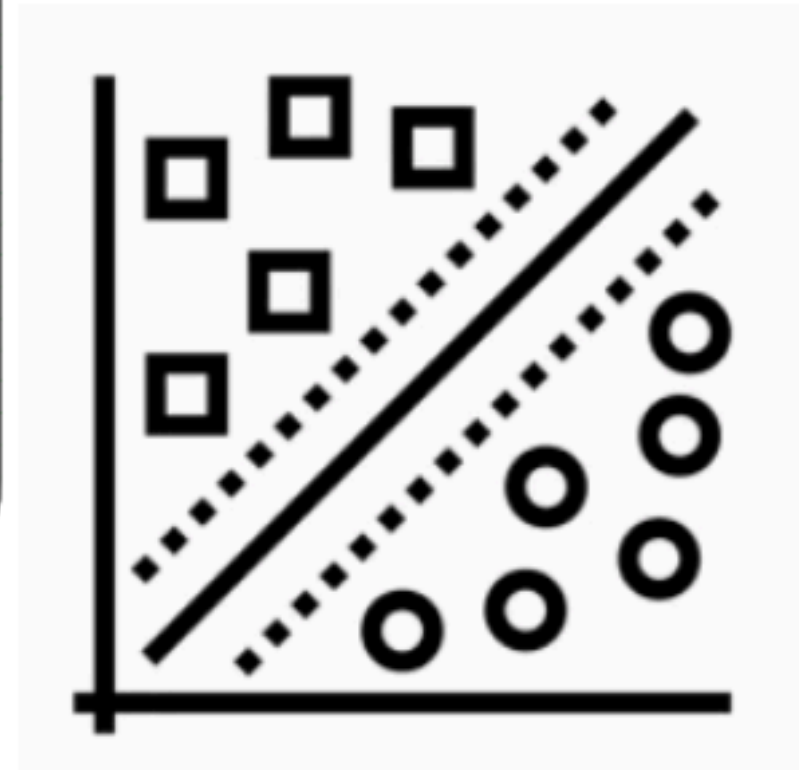
КЛАСИФІКАЦІЯ ЗА ДОПОМОГОЮ МЕТОДУ CART



Підготував: студент ОІ-32
Криворучко Микола

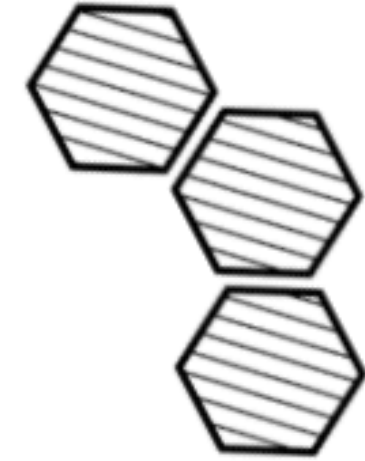
ЗМІСТ

1. Що таке Data Mining?
2. Які існують задачі Data Mining?
3. Задача класифікації
4. Методи класифікації
5. CART
6. Висновок



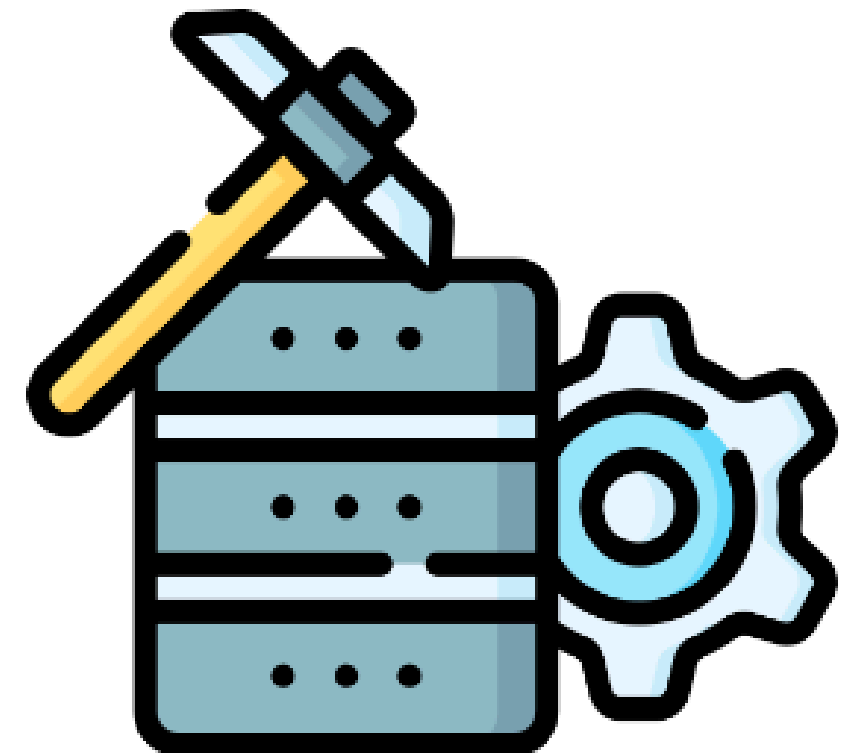
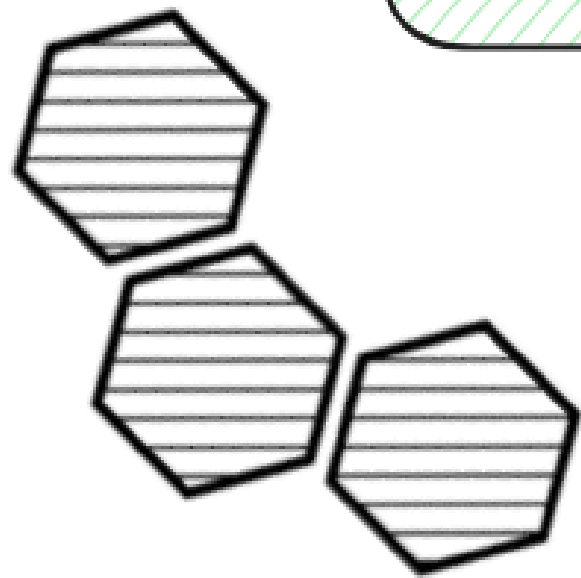
Що таке Data Mining?

Набір методів, алгоритмів та засобів
опрацювання "сирих даних" із метою
видобування з них необхідної
інформації(знань)



Задачі Data Mining

1. Задача класифікації
2. Задача регресії
3. Задача кластеризації
4. Побудова асоціативних правил



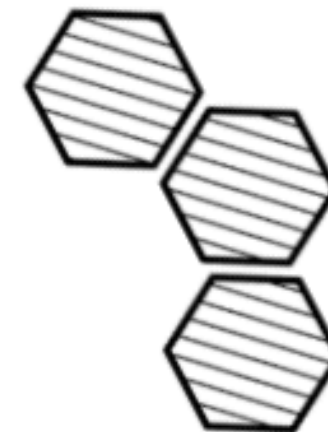
Задача класифікації

задача класифікації зводиться до визначення класу об'єкта по його характеристикам. Необхідно зауважити, що в цьому завданні множина класів, до яких може бути віднесений об'єкт, відомо заздалегідь.



Методи класифікації

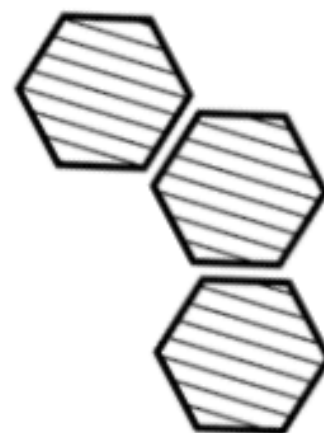
- ◆ метод найближчого сусіда
- ◆ наївна байєосва класифікація
- ◆ CBR
- ◆ група присвячених побудові
дерев рішень
- ◆ нейронні мережі
- ◆ генетичні алгоритми
- ◆ метод опорних векторів



CART

це алгоритм побудови бінарного дерева рішень, який може використовуватися:

- для класифікації (розбиття даних на категорії);
- для регресії (прогнозування числових значень).

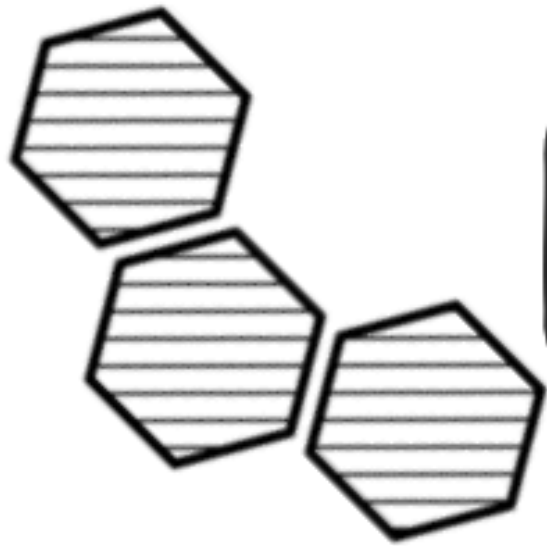
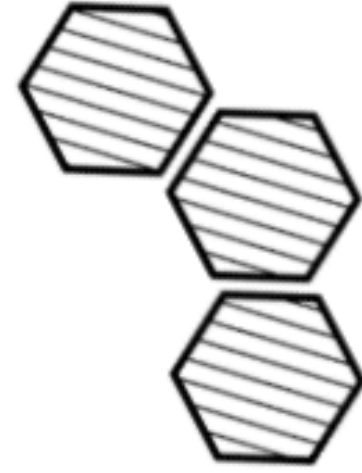


Основні принципи роботи

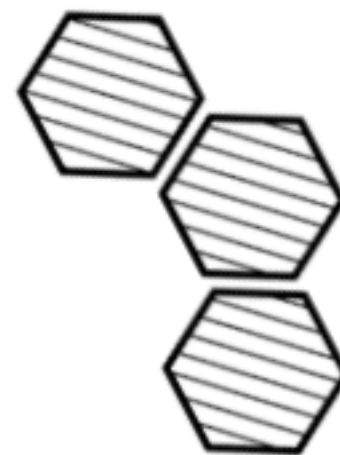
Бінарна структура: Кожен вузол дерева ділить вибірку на дві частини (ліву й праву гілку). Тому дерево CART завжди є бінарним (тобто кожен вузол має максимум двох нащадків).

Логіка розбиття: На кожному кроці алгоритм вибирає одну ознаку і один поріг, який найкраще ділить дані.

Мета — зробити підмножини максимально «чистими», тобто щоб приклади одного класу потрапляли в одну гілку.



Індекс Gini



Індекс Gini у CART

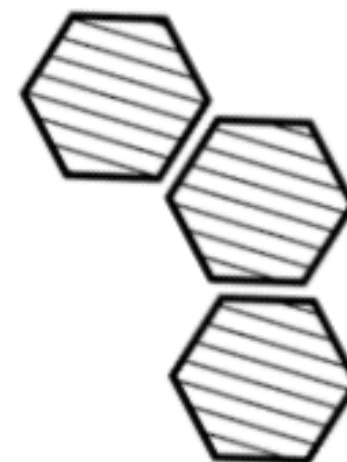
- Gini вузла: $Gini(t) = 1 - \sum_{i=1}^C p_i^2$
 - p_i — частка прикладів класу i у вузлі t
 - $0 \rightarrow$ вузол чистий (усі приклади одного класу)
 - $\max \rightarrow$ приклади рівномірно змішані
- Gini розбиття:

$$Gini_{split} = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2)$$



- N_1, N_2 — кількість прикладів у підвузлах
- Мінімізуємо $Gini_{split} \rightarrow$ найчистіше розбиття

Типи ознак



- Числові атрибути:

Правило виду: $x_i \leq c$,

де c — порогове значення (зазвичай середнє між двома сусідніми значеннями в даних).

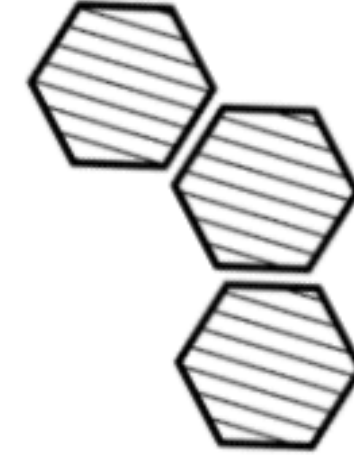
- Категоріальні атрибути:

Правило виду: $x_i \in V(x_i)$,

де $V(x_i)$ — підмножина можливих категорій.



Відсікання дерева



CART використовує метод Minimal Cost-Complexity Pruning

- Спочатку вирощуємо повне дерево, поки можливо.
- Потім поступово відсікаємо (скорочуємо) гілки, які мало покращують якість класифікації, щоб уникнути перенавчання.

Ідея: знайти баланс між точністю і простотою дерева.



V-fold Cross-Validation



Перехресна перевірка (V-fold Cross-Validation) використовується для вибору оптимального розміру дерева

- Дані діляться на V частин.
- Кожного разу дерево тренується на $V-1$ частинах і перевіряється на решті.
- Це дозволяє стабільно оцінити якість дерева, навіть при невеликому об'ємі даних.



CART

Чи заб'є Роналду гол в матчі?

	Сила суперника	Місце матчу	Форма Роналдо	Підтримка команди	Забив
0	Сильний	В гостях	Добра	Висока	Ні
1	Слабкий	Дома	Добра	Висока	Так
2	Середній	В гостях	Погана	Низька	Ні
3	Слабкий	Дома	Добра	Середня	Так
4	Сильний	Дома	Добра	Висока	Так
5	Середній	В гостях	Погана	Низька	Ні
6	Слабкий	В гостях	Добра	Висока	Так
7	Середній	Дома	Погана	Середня	Ні
8	Сильний	Дома	Добра	Середня	Так
9	Слабкий	В гостях	Погана	Висока	Ні
10	Середній	Дома	Добра	Висока	Так
11	Сильний	В гостях	Погана	Середня	Ні

CART

Сила суперника

Забив Сила суперника	Так	Ні	Всього
Слабкий	4	4	8
Сильний	2	2	4

$$Gini(\text{сила суперника}=\text{слабкий}) = 1 - (4/8)^2 - (4/8)^2 = 0.5$$

$$Gini(\text{сила суперника}=\text{сильний}) = 1 - (2/4)^2 - (2/4)^2 = 0.5$$

$$Gini(\text{сила суперника}) = 8/12 * 0.5 + 4/12 * 0.5 = 0.5$$

CART

Місце матчу

Забив Місце матчу	Так	Ні	Всього
В гостях	5	1	6
Дома	1	5	6

$$\text{Gini}(\text{місце матчу}=\text{в гостях}) = 1 - (5/6)^2 - (1/6)^2 = 0.2778$$

$$\text{Gini}(\text{місце матчу}=\text{дома}) = 1 - (1/6)^2 - (5/6)^2 = 0.2778$$

$$\text{Gini}(\text{місце матчу}) = 8/12 * 0.2778 + 4/12 * 0.2778 = 0.2778$$

CART

Форма Роналду

Форма \ Забив	Так	Ні	Всього
Добра	1	6	7
Погана	5	0	5

$$Gini(\text{форма Роналду}=\text{добра}) = 1 - (1/7)^2 - (6/7)^2 = 0.2449$$

$$Gini(\text{форма Роналду}=\text{погана}) = 1 - (5/5)^2 - (0/5)^2 = 0$$

$$Gini(\text{форма Роналду}) = 7/12 * 0.2449 + 5/12 * 0 = 0.1429$$

CART

Підтримка команди

Забив Підтримка	Так	Ні	Всього
Висока	2	4	6
Низька	4	2	6

$$\text{Gini}(\text{місце матчу}=\text{в гостях}) = 1 - (2/6)^2 - (4/6)^2 = 0.444$$

$$\text{Gini}(\text{місце матчу}=\text{дома}) = 1 - (4/6)^2 - (2/6)^2 = 0.444$$

$$\text{Gini}(\text{місце матчу}) = 6/12 * 0.444 + 6/12 * 0.444 = 0.444$$

CART

Ознака

Індекс Gini

Місце матчу

0.2778

Форма Роналду

0.1429

Підтримка команди

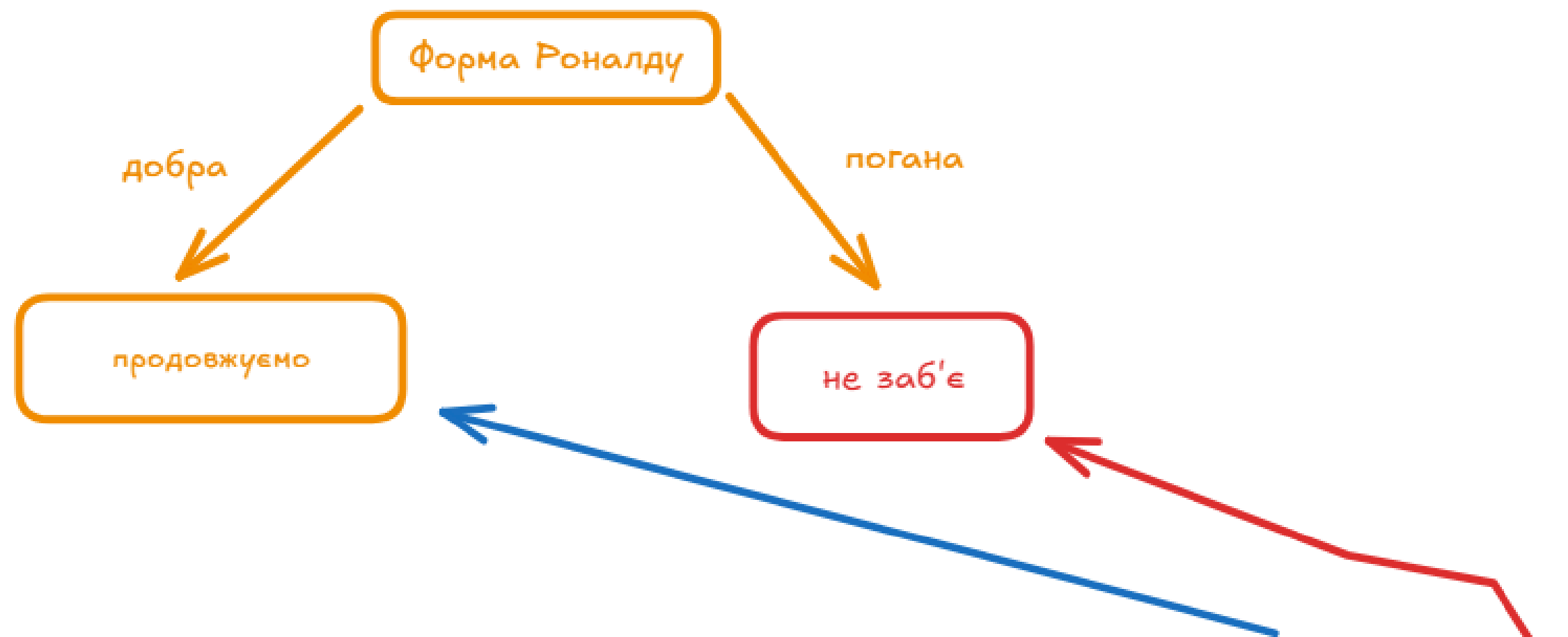
0.444

Сила суперника

0.5



CART



$$Gini(\text{форма Роналду}=\text{добра}) = 1 - (1/7)^2 - (6/7)^2 = 0.2449$$

$$Gini(\text{форма Роналду}=\text{погана}) = 1 - (5/5)^2 - (0/5)^2 = 0$$

$$Gini(\text{форма Роналду}) = 7/12 * 0.2449 + 5/12 * 0 = 0.1429$$

CART

	Сила суперника	Місце матчу	Форма Роналдо	Підтримка команди	Забив
0	Сильний	В гостях	Добра		Ні
1	Слабкий	Дома	Добра		Так
3	Слабкий	Дома	Добра		Так
4	Сильний	Дома	Добра		Так
6	Слабкий	В гостях	Добра		Так
8	Сильний	Дома	Добра		Так
10	Слабкий	Дома	Добра		Так

CART

Сила суперника:

$$Gini(\text{Сильный}) = 1 - (1/3)^2 - (2/3)^2 = 0.4444$$

$$Gini(\text{Слабкий}) = 1 - (4/4)^2 = 0.0000$$

$$Gini(\text{Сила суперника}) = 3/7 \times 0.4444 + 4/7 \times 0.0000 = \underline{0.1905}$$

Місце матчу:

$$Gini(\text{В гостях}) = 1 - (1/2)^2 - (1/2)^2 = 0.5000$$

$$Gini(\text{Дома}) = 1 - (5/5)^2 = 0.0000$$

$$Gini(\text{Місце матчу}) = 2/7 \times 0.5000 + 5/7 \times 0.0000 = \underline{0.1429} \leftarrow \text{Найменший Gini}$$

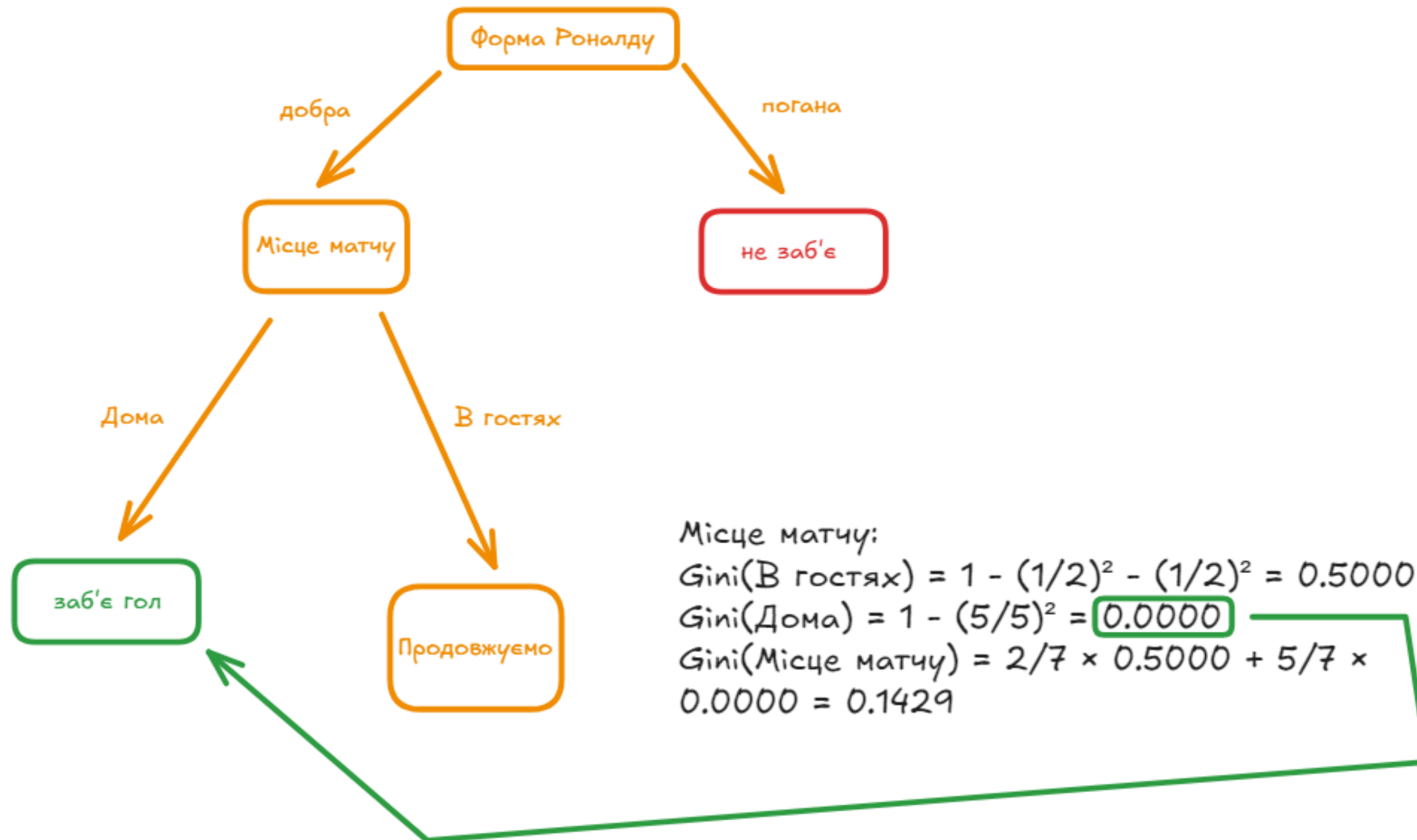
Підтримка команди:

$$Gini(\text{Висока}) = 1 - (1/5)^2 - (4/5)^2 = 0.3200$$

$$Gini(\text{Низька}) = 1 - (2/2)^2 = 0.0000$$

$$Gini(\text{Підтримка команди}) = 5/7 \times 0.3200 + 2/7 \times 0.0000 = \underline{0.2286}$$

CART



CART

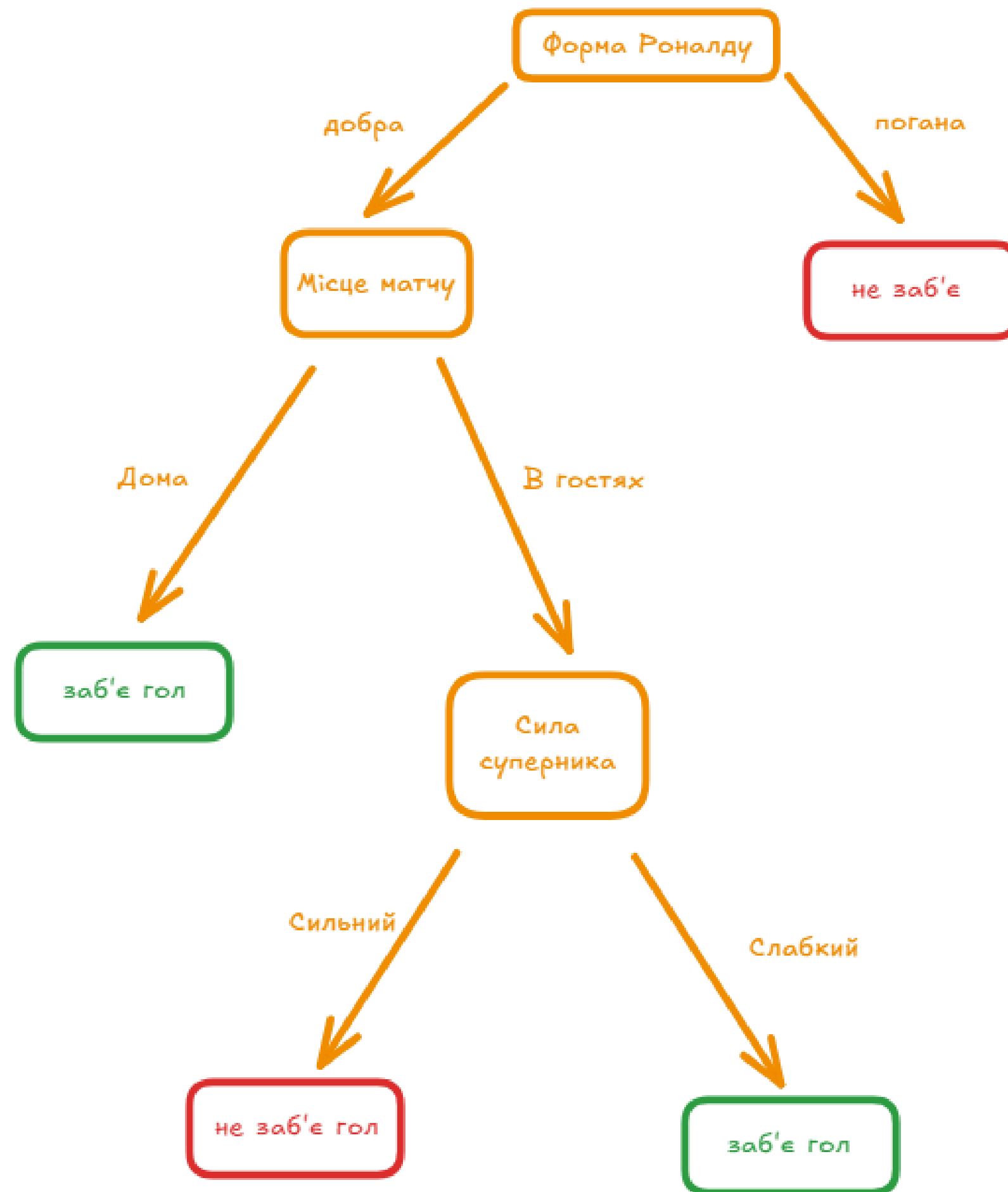
	Сила суперника	Місце матчу	Форма Роналдо	Підтримка команди	Забив
0	Сильний	В гостях	Добра	Висока	Ні
6	Слабкий	В гостях	Добра	Висока	Так

Сила суперника:

$$Gini(\text{Сильний}) = 1 - (1/1)^2 = 0.0000$$

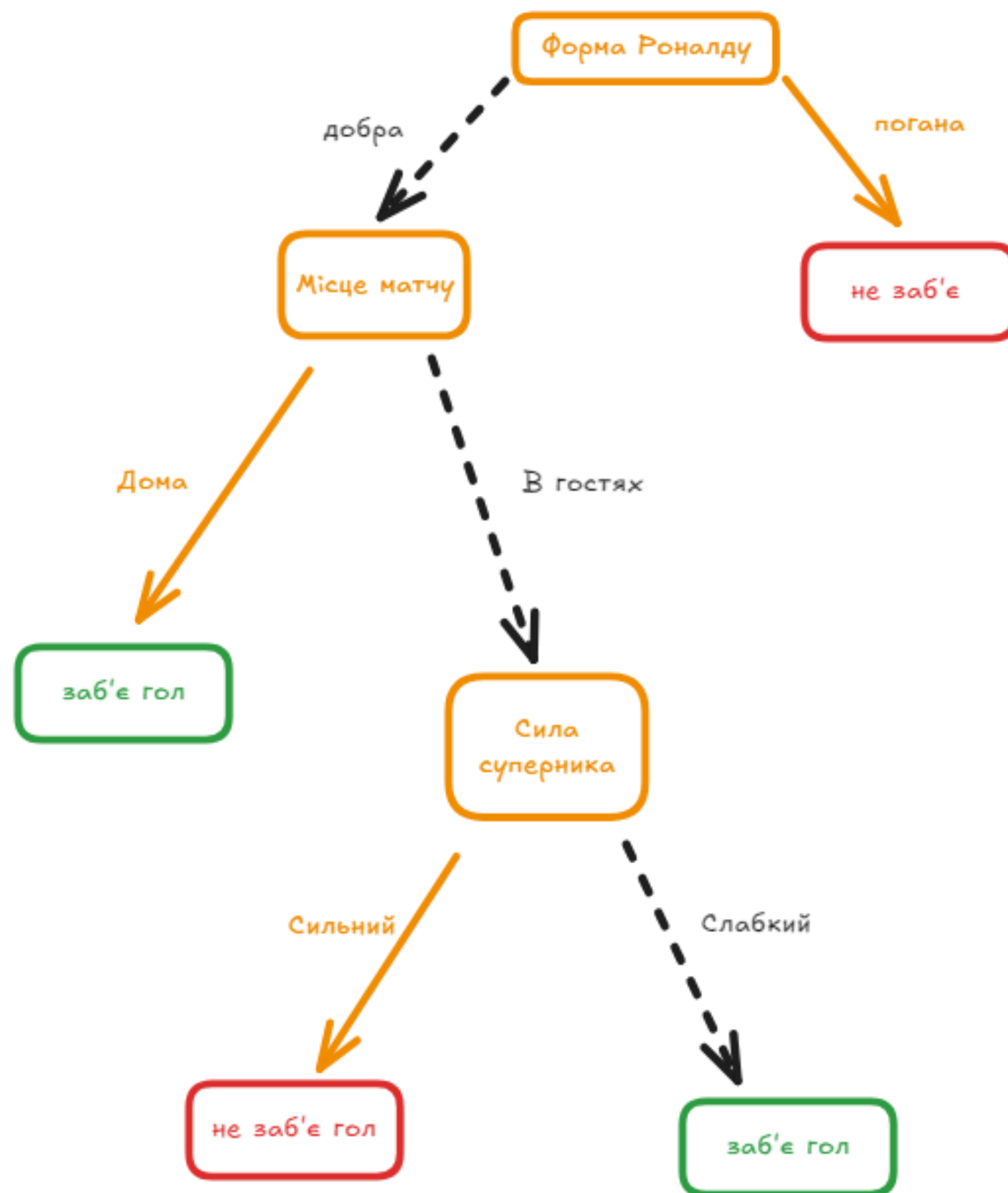
$$Gini(\text{Слабкий}) = 1 - (1/1)^2 = 0.0000$$

$$Gini(\text{Сила суперника}) = 1/2 \times 0.0000 + 1/2 \times 0.0000 = 0.0000$$



Тестові дані

Форма Роналду: Добра
Місце матчу: В гостях
Сила суперника: Слабкий
Підтримка команди: Висока



Висновок

У цій роботі я реалізував алгоритм CART для побудови дерева рішень. Я розробив функції для обчислення індексу Gini та автоматичного пошуку оптимальних розбиттів.