Інтелектуальний аналіз даних

# АЛГОРИТМ ID3

Виконав:студент групи ОІ-36 Пироженко Назар

# ЗМІСТ

# ЩО ТАКЕ ДЕРЕВА РІШЕНЬ?

**Дерева рішень** - це ієрархічні структури, що використовуються для класифікації та прогнозування на основі набору правил типу «Якщо…, то…». Кожен внутрішній вузол дерева відповідає перевірці певної ознаки, кожна гілка - результату перевірки, а кожен лист - значенню цільової змінної (класу або числового значення).

# ОСНОВНІ ПОНЯТТЯ

- Корінь — початковий вузол дерева.
- Внутрішній вузол — містить умову для розгалуження.
- Лист — кінцевий вузол, що визначає результат.
- Підтримка — кількість прикладів, які класифікуються вузлом.
- Достовірність — частка правильно класифікованих прикладів у вузлі.

# ЩО ТАКЕ ID3?

**ID3** - це алгоритм, який використовується для генерації дерев рішень у машинному навчанні з деякого набору даних. ID3 є попередником алгоритму C4.5.

**Алгоритм ID3** не гарантує знаходження оптимального рішення. Він використовує жадібну стратегію, обираючи локальний кращий атрибут для розбиття множини даних на кожній ітерації. Оптимальність алгоритму може бути покращена шляхом використання пошуку з вертанням під час пошуку оптимального дерева рішень, але це може призвести до погіршення швидкості роботи.

# ОСНОВНІ КРОКИ АЛГОРИТМУ:

- Обчислити ентропію загальної множини S;
- Обчислити інформаційні прирости кожного з атрибутів множини S;
- Коренем дерева стане атрибут з найбільшим інформаційним приростом;
- Повторити попередні кроки для всіх підмножин, використовуючи атрибути, що залишились.

# ФОРМУЛИ, ЯКІ ВИКОРИСТОВУЮТЬСЯ ПРИ РЕАЛІЗАЦІЇ ДАНОГО АЛГОРИТМУ:

$$H(S) = -\sum_{i=0}^{i=n} p(x_i)\log_2 p(x_i)$$ – ентропія множини $S$

$IG(S,A) = H(S) - H(S|A)$ – інформаційний приріст множини $S$

при розділенні атрибутом A

$$H(S|A) = \sum_{t \in T} p(t)H(t)$$ – ентропія множини $S$

після розбиття за атрибутом A

## ПЕРЕВАГИ АЛГОРИТМУ

- Легко інтерпретується (через простоту моделі,

можна легко відобразити дерево і простежити за всіма вузлами дерева).

- Простота класифікації.

## НЕДОЛІКИ АЛГОРИТМУ

- ·Алгоритм не працює з неперервними (числовими) атрибутами.
- ·Алгоритм не передбачає можливості роботи з пропусками в даних.

# НАВЧАЛЬНА ВИБІРКА

Об’єкт дослідження: доречність прогулянки людини

Для кожного атрибуту ми знаємо всі можливі значення:
**Прогноз**: {Сонячно, Похмуро, Дощ}
**Температура**: {Гаряча, М’яка, Прохолодна}
**Вологість**: {Висока, Нормальна}
**Вітер**: {Слабкий, Сильний}

| День | Прогноз | Температура | Вологість | Вітер | Прогулянка |
|---|---|---|---|---|---|
| D1 | Сонячно | Гаряча | Висока | Слабкий | Так |
| D2 | Сонячно | Гаряча | Висока | Сильний | Ні |
| D3 | Сонячно | М'яка | Висока | Слабкий | Так |
| D4 | Сонячно | Прохолодна | Нормальна | Слабкий | Так |
| D5 | Похмуро | Гаряча | Висока | Слабкий | Так |
| D6 | Похмуро | Гаряча | Висока | Сильний | Ні |
| D7 | Похмуро | М'яка | Висока | Слабкий | Так |
| D8 | Похмуро | Прохолодна | Нормальна | Сильний | Так |
| D9 | Дощ | М'яка | Висока | Слабкий | Ні |
| D10 | Дощ | М'яка | Висока | Сильний | Ні |
| D11 | Дощ | Прохолодна | Нормальна | Слабкий | Так |
| D12 | Дощ | Прохолодна | Нормальна | Сильний | Ні |
| D13 | Сонячно | М'яка | Нормальна | Слабкий | Так |
| D14 | Похмуро | М'яка | Нормальна | Сильний | Так |

# ОБРАХУНОК ЕНТРОПІЇ ВСІЄЇ ВИБІРКИ

- Цільовий атрибут — "Прогулянка". У нас є 14 днів.
- Так (Y): 9 днів (D1, D3, D4, D5, D7, D8, D11, D13, D14)
- Ні (N): 5 днів (D2, D6, D9, D10, D12)
- Формула ентропії: $H(S) = -p_+ * \log2(p_+) - p_- * \log2(p_-)$
- $p_+$ (ймовірність "Так") = 9/14
- $p_-$ (ймовірність "Ні") = 5/14
- $H(S) = - (9/14) * \log2(9/14) - (5/14) * \log2(5/14)$
- $H(S) \approx - (0.642) * (-0.637) - (0.357) * (-1.486)$
- $H(S) \approx 0.409 + 0.531$
- $H(S) \approx 0.940$
- Ентропія вихідної вибірки $H(S) \approx 0.940$

# ОБРАХУНОК ІНФОРМАЦІЙНОГО ПРИРОСТУ ДЛЯ КОЖНОГО АТРИБУТУ ОКРЕМО

**Атрибут "Прогноз":**

- Сонячно: 5 днів (4 Так, 1 Ні) | H = 0.722
- Похмуро: 5 днів (4 Так, 1 Ні) | H = 0.722
- Дощ: 4 дні (1 Так, 3 Ні) | H = 0.811

IG(Прогноз) = 0.940 - [(5/14)*0.722 + (5/14)*0.722 + (4/14)*0.811] = 0.193

**Атрибут "Температура":**

- Гаряча: 4 дні (2 Так, 2 Ні) | H = 1.0
- М'яка: 6 днів (4 Так, 2 Ні) | H = 0.918
- Прохолодна: 4 дні (3 Так, 1 Ні) | H = 0.811

IG(Температура) = 0.940 - [(4/14)*1.0 + (6/14)*0.918 + (4/14)*0.811] = 0.029

**Атрибут "Вологість":**

- Висока: 8 днів (4 Так, 4 Ні) | H = 1.0
- Нормальна: 6 днів (5 Так, 1 Ні) | H = 0.650

IG(Вологість) = 0.940 - [(8/14)*1.0 + (6/14)*0.650] = 0.151

**Атрибут "Вітер":**

- Слабкий: 8 днів (7 Так, 1 Ні) | H = 0.544
- Сильний: 6 днів (2 Так, 4 Ні) | H = 0.918

IG(Вітер) = 0.940 - [(8/14)*0.544 + (6/14)*0.918] = 0.048

## СОРТУВАННЯ ІНФОРМАЦІЙНИХ ПРИРОСТІВ КОЖНОГО З АТРИБУТІВ ЗА СПАДАННЯМ

1. Прогноз: 0.193
2. Вологість: 0.151
3. Вітер: 0.048
4. Температура: 0.029

## ОБРАННЯ КОРЕНЮ ДЕРЕВА

Коренем дерева стає атрибут з найвищим інформаційним приростом — "Прогноз".

# РОЗДІЛЕННЯ ЗА "ПРОГНОЗ"

Гілка A: Прогноз = Сонячно (5 днів: D1,D2,D3,D4,D13)
Розподіл: 4 Так, 1 Ні | H = 0.722

Гілка B: Прогноз = Похмуро (5 днів: D5,D6,D7,D8,D14)
Розподіл: 4 Так, 1 Ні | H = 0.722

Гілка C: Прогноз = Дощ (4 дні: D9,D10,D11,D12)
Розподіл: 1 Так, 3 Ні | H = 0.811

# ОБРОБКА ГІЛКИ "СОНЯЧНО"

Температура:

- Гаряча (D1,D2): 1 Так, 1 Ні | H = 1.0
- М'яка (D3,D13): 2 Так, 0 Ні | H = 0
- Прохолодна (D4): 1 Так, 0 Ні | H = 0
- IG = 0.722 - [(2/5)*1.0 + (2/5)*0 + (1/5)*0] = 0.322

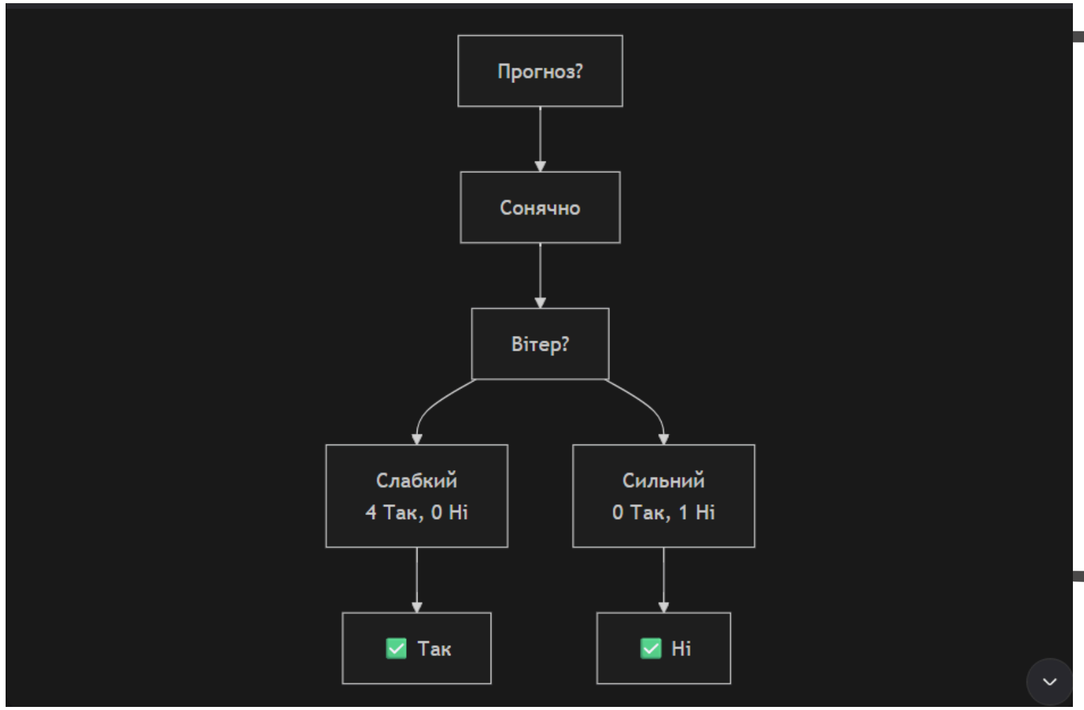
Вологість:

- Висока (D1,D2,D3): 2 Так, 1 Ні | H = 0.918
- Нормальна (D4,D13): 2 Так, 0 Ні | H = 0
- IG = 0.722 - [(3/5)*0.918 + (2/5)*0] = 0.171

Вітер:

- Слабкий (D1,D3,D4,D13): 4 Так, 0 Ні | H = 0
- Сильний (D2): 0 Так, 1 Ні | H = 0
- IG = 0.722 - [(4/5)*0 + (1/5)*0] = 0.722

MAX IG = 0.722 (Вітер)



# ОБРОБКА ГІЛКИ "ПОХМУРО"

Температура:

- Гаряча (D5,D6): 1 Так, 1 Ні | H = 1.0
- М'яка (D7,D14): 2 Так, 0 Ні | H = 0
- Прохолодна (D8): 1 Так, 0 Ні | H = 0
- IG = 0.722 - [(2/5)*1.0 + (2/5)*0 + (1/5)*0] = 0.322

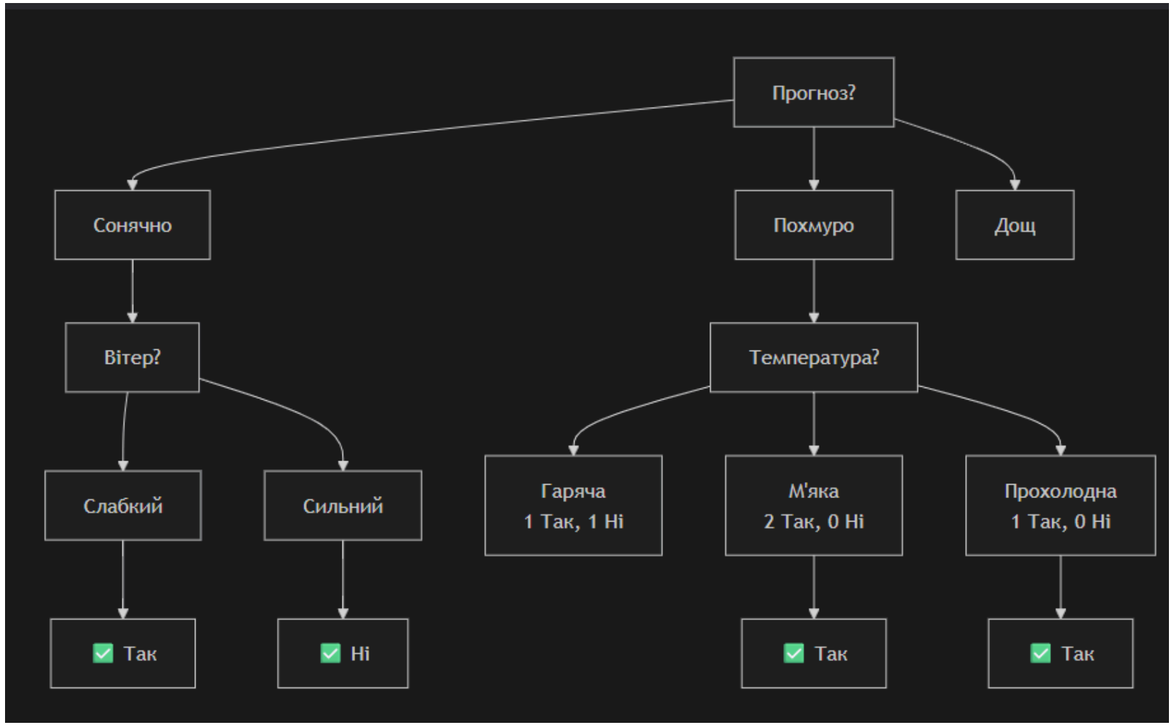
Вологість:

- Висока (D5,D6,D7): 2 Так, 1 Ні | H = 0.918
- Нормальна (D8,D14): 2 Так, 0 Ні | H = 0
- IG = 0.722 - [(3/5)*0.918 + (2/5)*0] = 0.171

Вітер:

- Слабкий (D5,D7): 2 Так, 0 Ні | H = 0
- Сильний (D6,D8,D14): 2 Так, 1 Ні | H = 0.918
- IG = 0.722 - [(2/5)*0 + (3/5)*0.918] = 0.171
-

MAX IG = 0.322 (Температура)



# ОБРОБКА ГІЛКИ "ПОХМУРО > ГАРЯЧА"

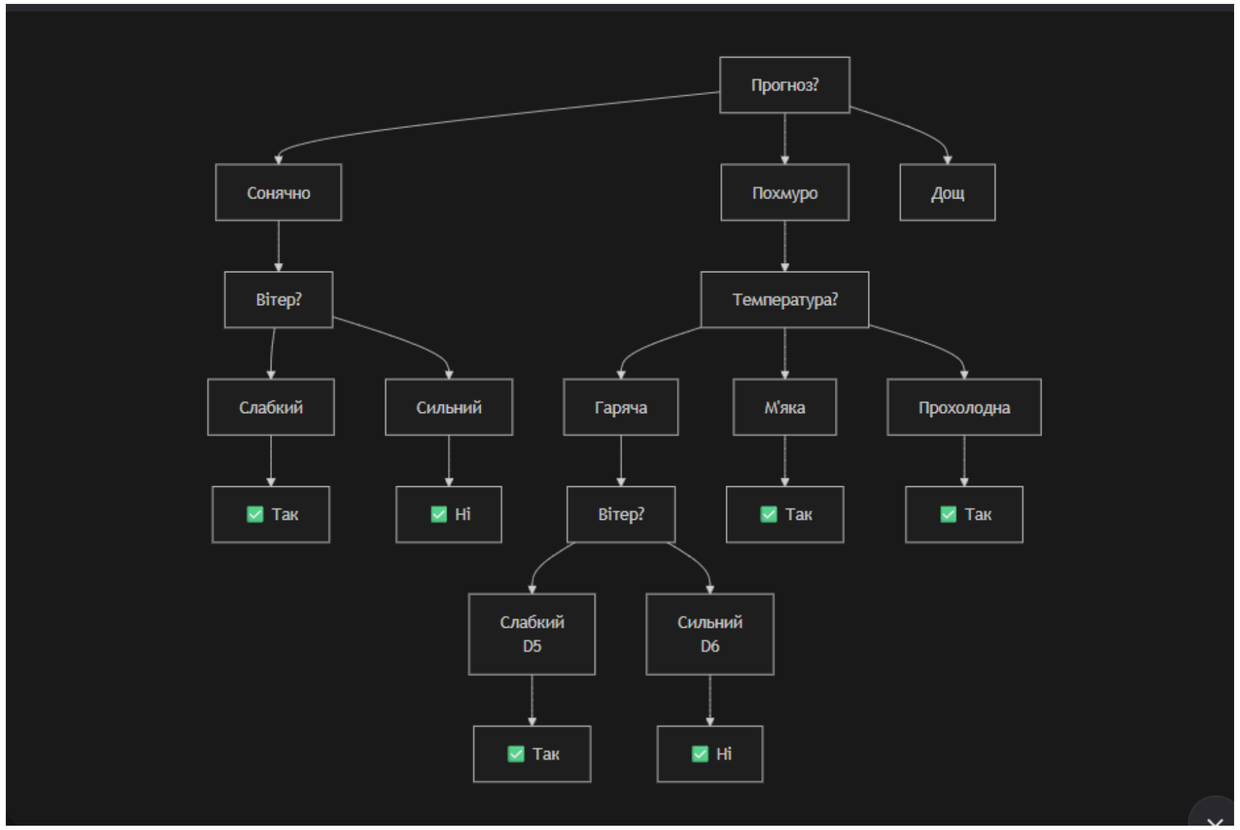
Дані: D5,D6 (1 Так, 1 Ні) | H = 1.0
Залишилися атрибути: Вологість, Вітер
IG для Вологість:

- Висока (D5,D6): 1 Так, 1 Ні | H = 1.0
- IG = 1.0 - [(2/2)*1.0] = 0

IG для Вітер:

- Слабкий (D5): 1 Так, 0 Ні | H = 0
- Сильний (D6): 0 Так, 1 Ні | H = 0
- IG = 1.0 - [(1/2)*0 + (1/2)*0] = 1.0

MAX IG = 1.0 (Вітер)



# ОБРОБКА ГІЛКИ "ДОЩ"

Температура:

- М'яка (D9,D10): 0 Так, 2 Ні | H = 0
- Прохолодна (D11,D12): 1 Так, 1 Ні | H = 1.0
- IG = 0.811 - [(2/4)*0 + (2/4)*1.0] = 0.311

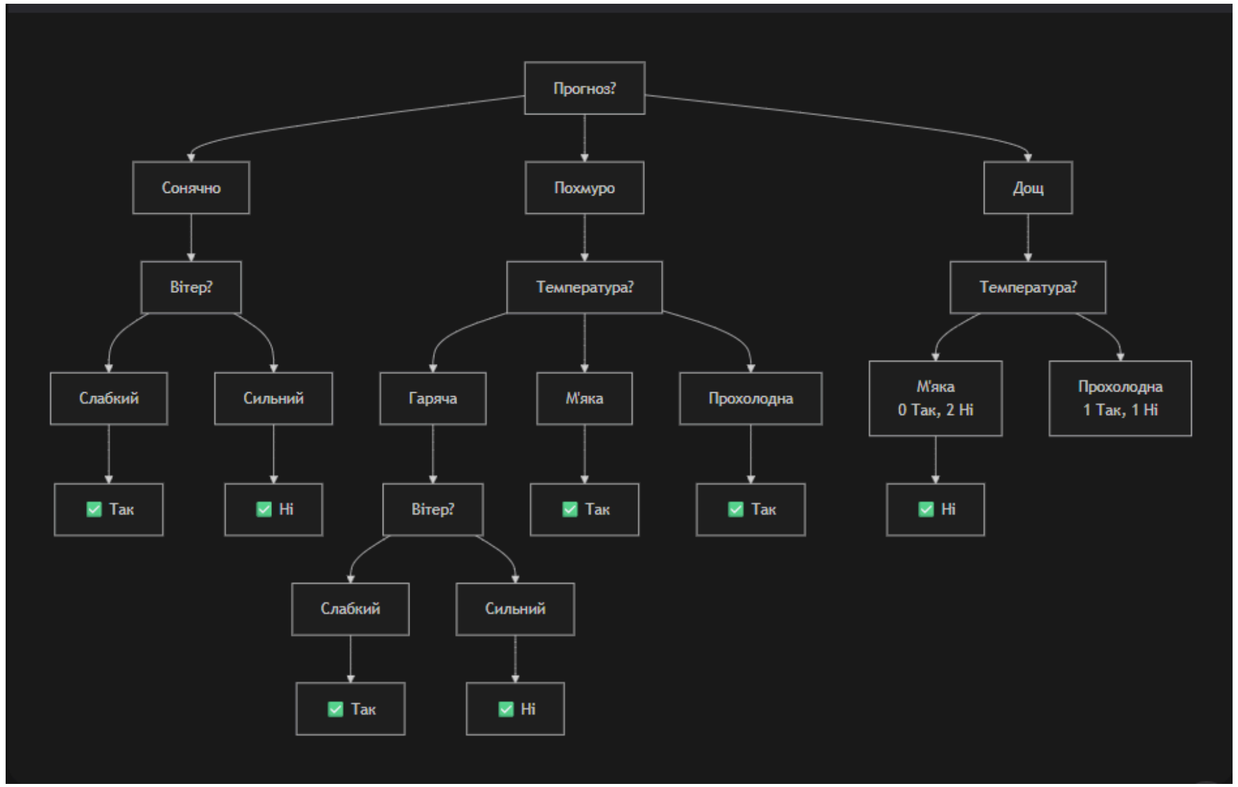
Вологість:

- Висока (D9,D10): 0 Так, 2 Ні | H = 0
- Нормальна (D11,D12): 1 Так, 1 Ні | H = 1.0
- IG = 0.811 - [(2/4)*0 + (2/4)*1.0] = 0.311

Вітер:

- Слабкий (D9,D11): 1 Так, 1 Ні | H = 1.0
- Сильний (D10,D12): 0 Так, 2 Ні | H = 0
- IG = 0.811 - [(2/4)*1.0 + (2/4)*0] = 0.311

MAX IG = 0.311 (обидва однакові, обираємо Температура)



# ОБРОБКА ГІЛКИ "ДОЩ > ПРОХОЛОДНА"

Дані: D11,D12 (1 Так, 1 Ні) | H = 1.0

Залишилися атрибути: Вологість, Вітер

IG для Вологість:

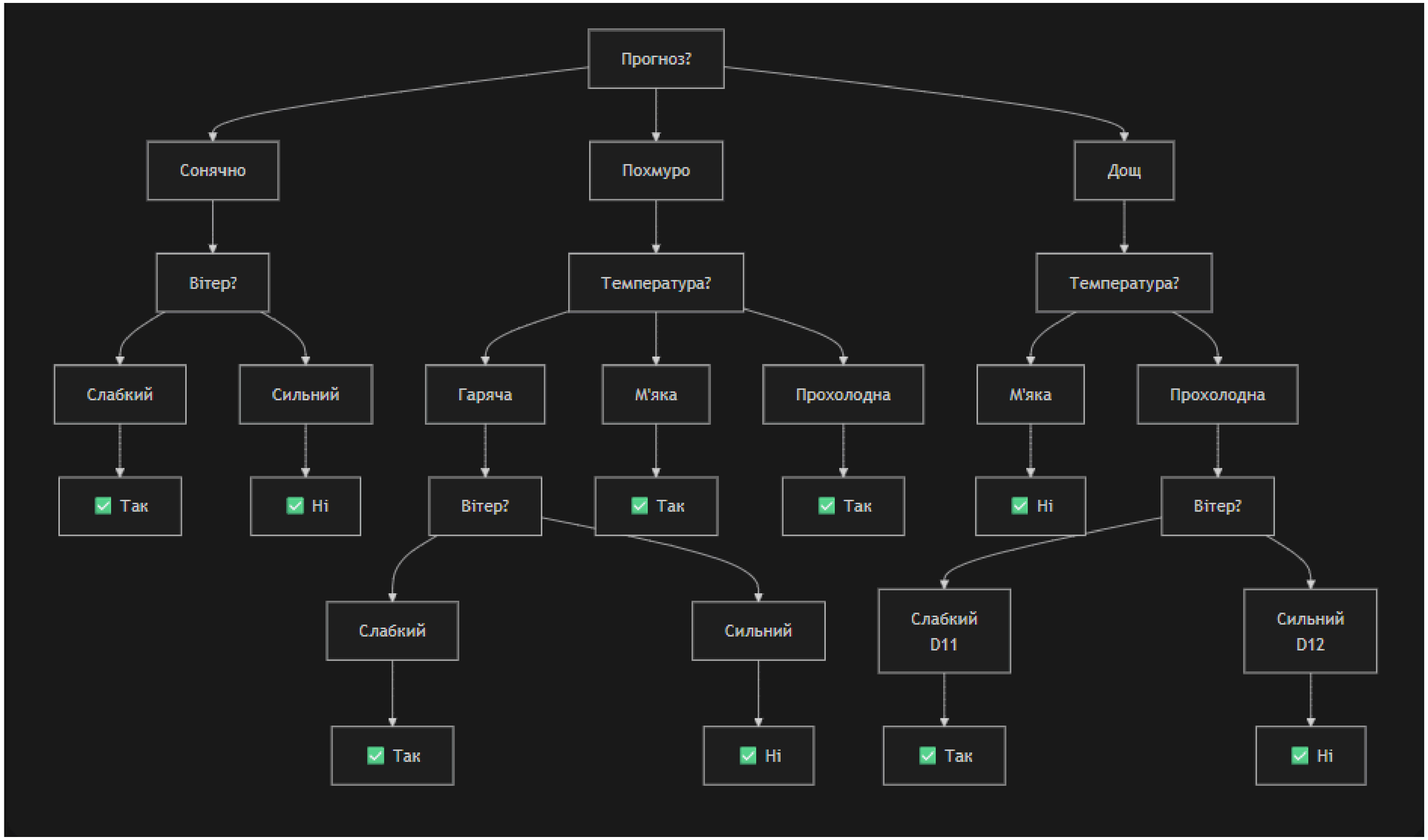
- Нормальна (D11,D12): 1 Так, 1 Ні | H = 1.0
- IG = 1.0 - [(2/2)*1.0] = 0

IG для Вітер:

- Слабкий (D11): 1 Так, 0 Ні | H = 0
- Сильний (D12): 0 Так, 1 Ні | H = 0
- IG = 1.0 - [(1/2)*0 + (1/2)*0] = 1.0

MAX IG = 1.0 (Вітер)

# ФІНАЛЬНЕ ДЕРЕВО



# ВИСНОВОК

**Метод ID3** для побудови дерев рішень є одним з фундаментальних алгоритмів машинного навчання, що демонструє ефективність у завданнях класифікації та прийняття рішень. Незважаючи на свою простоту, він знаходить застосування в різних галузях, включаючи медичну діагностику, фінансовий аналіз та підтримку прийняття рішень. Алгоритм забезпечує зрозумілу інтерпретованість результатів, що є ключовою перевагою порівняно з багатьма складними моделями.

У цій роботі я реалізував повний цикл побудови дерева рішень за допомогою методу ID3, починаючи від обчислення ентропії та інформаційного приросту, закінчуючи формуванням оптимальної структури дерева. Практично підтверджено, що алгоритм адаптивно обирає атрибути для розділення на кожному кроці, максимізуючи інформаційний приріст для конкретної підмножини даних. Отримане дерево забезпечує точну класифікацію всіх навчальних прикладів, демонструючи ефективність методу у вирішенні завдань прийняття рішень на основі даних.