

ПОРІВНЯННЯ ОПТИМАЛЬНОСТІ АЛГОРИТМІВ КЛАСИФІКАЦІЇ ДАНИХ НА ПРИКЛАДІ ДЕРЕВ РІШЕНЬ

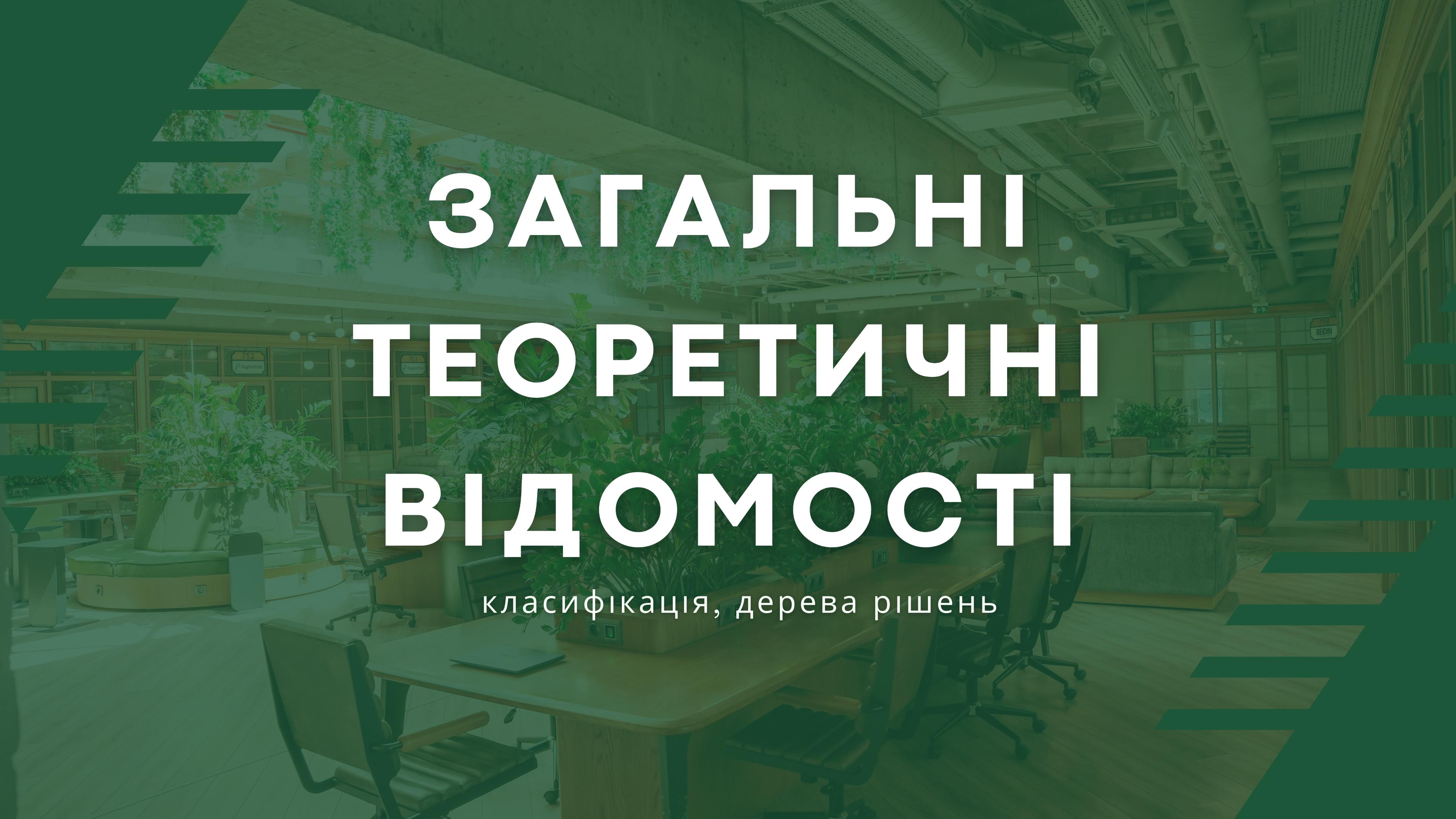
Бень Христина
Квасниця Галина



Зміст

- 01 Класифікація
- 02 Дерево рішень
- 03 Алгоритми побудови
- 04 Порівняння за часом
- 05 Таблиця спряженості
- 06 Порівняння за точністю
- 07 ROC - крива





ЗАГАЛЬНІ ТЕОРЕТИЧНІ ВІДОМОСТІ

класифікація, дерева рішень



1.

складаються з

- вузлів, які представляють ознаки,
- гілок, які вказують на можливі значення ознак,
- листів, які представляють результати класифікації або прогнозування.



2.

використовуються для прийняття рішень на основі
вхідних ознак, поділених на категорії



3.

працюють шляхом поділу даних на частини,
що розділяються по певним правилам.



4.

ієрархічна структура

Області використання

Класифікація



?



Регресія

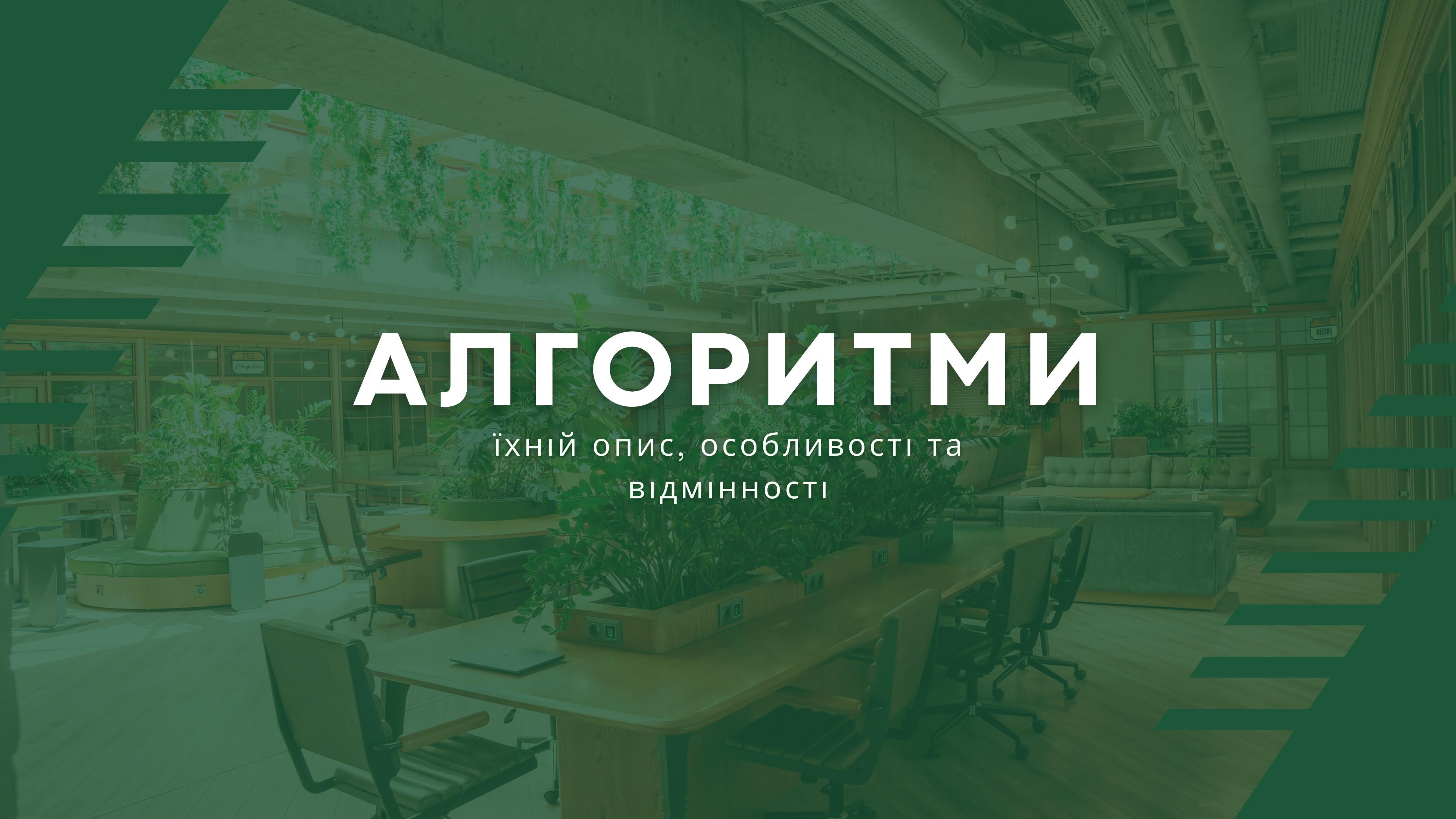
?



Prediction:
8 cm

розподіл об'єктів у певні класи

прогнозування неперервних значень



АЛГОРИТМИ

їхній опис, особливості та
відмінності



IDЗ



Критерій поділу

використовує **ентропію**
та інформаційну
вигоду для вибору
найкращого поділу



Недоліки

схильний до
перенавчання



Переваги

легкий у реалізації та
розумінні

C45



Критерій поділу

використовує **ентропію** та інформаційну **вигоду** для вибору найкращого поділу



Недоліки

складніший ніж ID3



Переваги

обробляє пропущені значення, виконує обрізання дерева для уникнення перенавчання

CART



Критерій поділу

використовує **Критерій
Джині**



Недоліки

чутливість до
невеликих змін у даних



Переваги

використовується для
класифікації та регресії

RANDOM FOREST



Критерій поділу

використовує кілька дерев з різними критеріями поділу



Недоліки

потребує більше обчислювальних ресурсів



Переваги

- висока стійкість до перенавчання,
- добра загальна точність

РЕЗУЛЬТАТИ

підрахунок показників
протестованих на різних наборах
даних

Час у секундах

25

125

4 000

id3

0.50800633

0.161302328

12.5053253173

C45

0.47323369

0.135096073

0.92654061317

CART

0.01834154

0.031525135

77.9678659439

Random forest

0.10077023

0.030347825

7.1440057754

Таблиця спряженості (Confusion Matrix)

► **TP (True Positives)**

правильно класифіковані зразки (істинно-позитивні випадки);

► **TN (True Negatives)**

правильно класифіковані негативні приклади (істинно-негативні випадки);

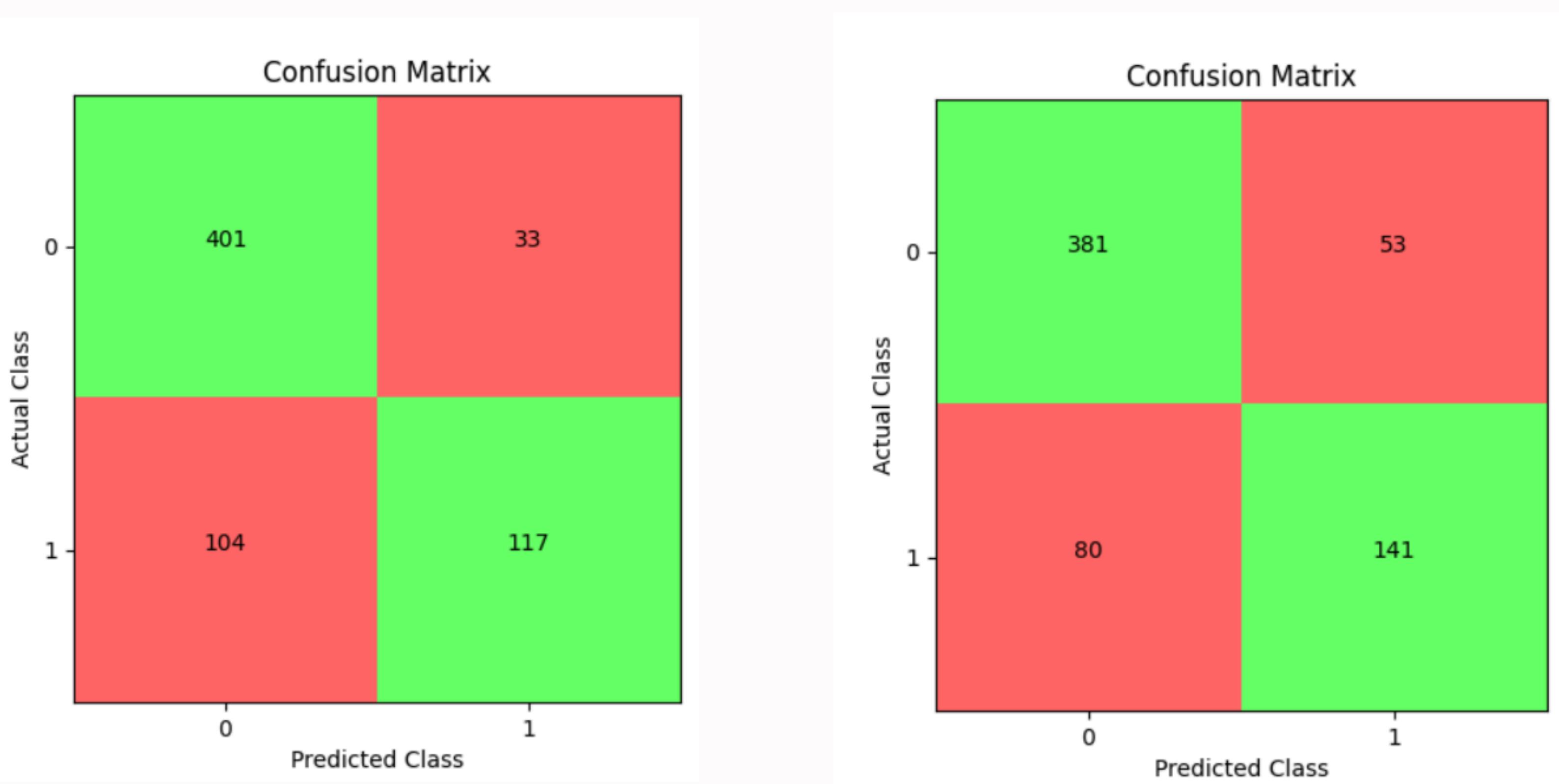
► **FN (False Negatives)**

позитивні зразки, класифіковані як негативні. Це помилка 1-го роду (хибно-негативні зразки);

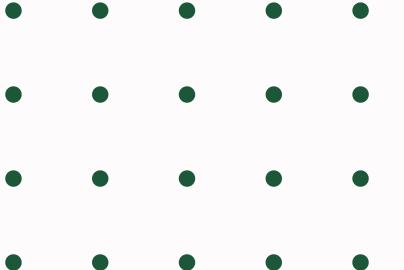
► **FP (False Positives)**

негативні зразки, класифіковані як позитивні. Це помилка 2-го роду (хибно-позитивні випадки)

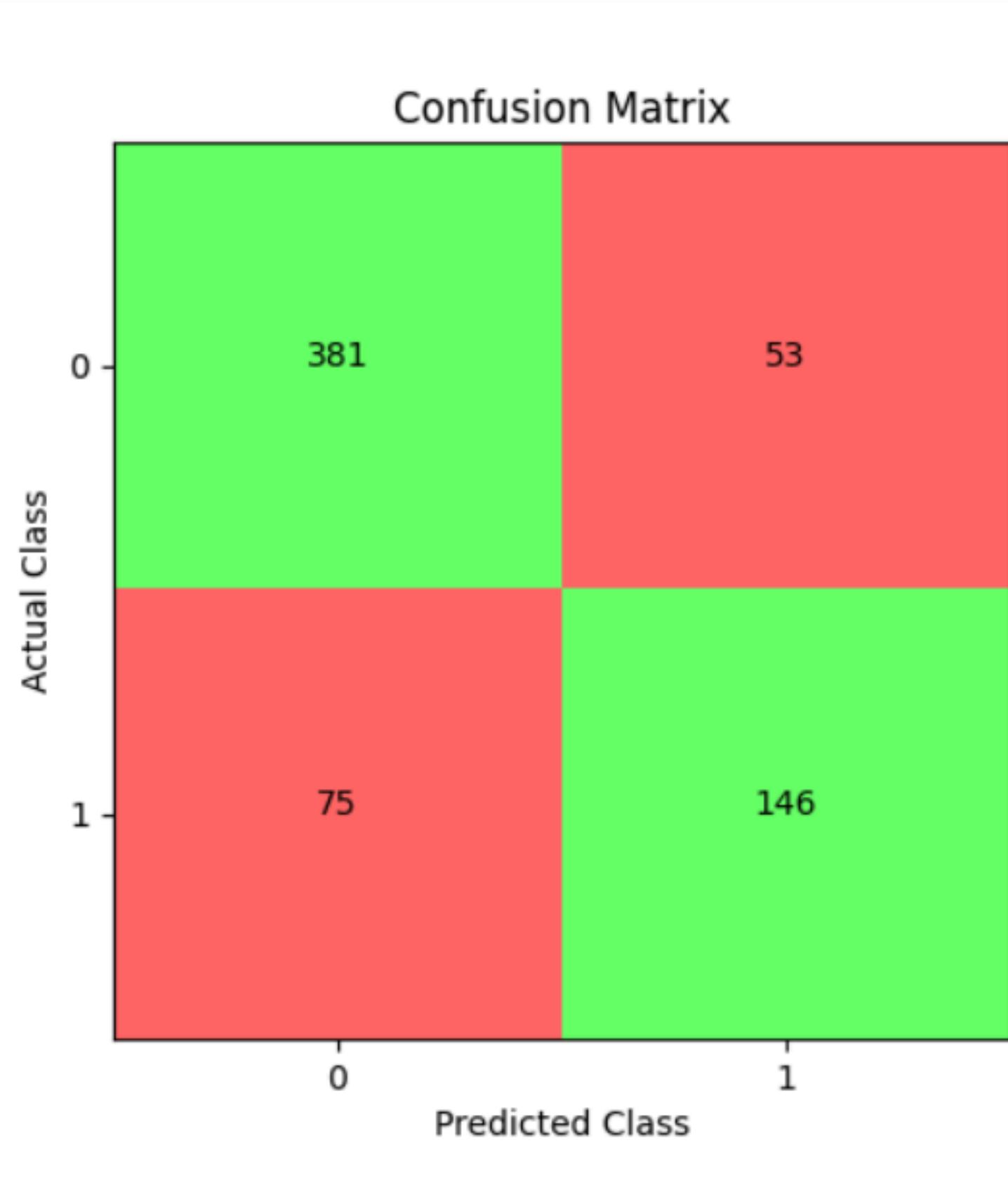
ID3



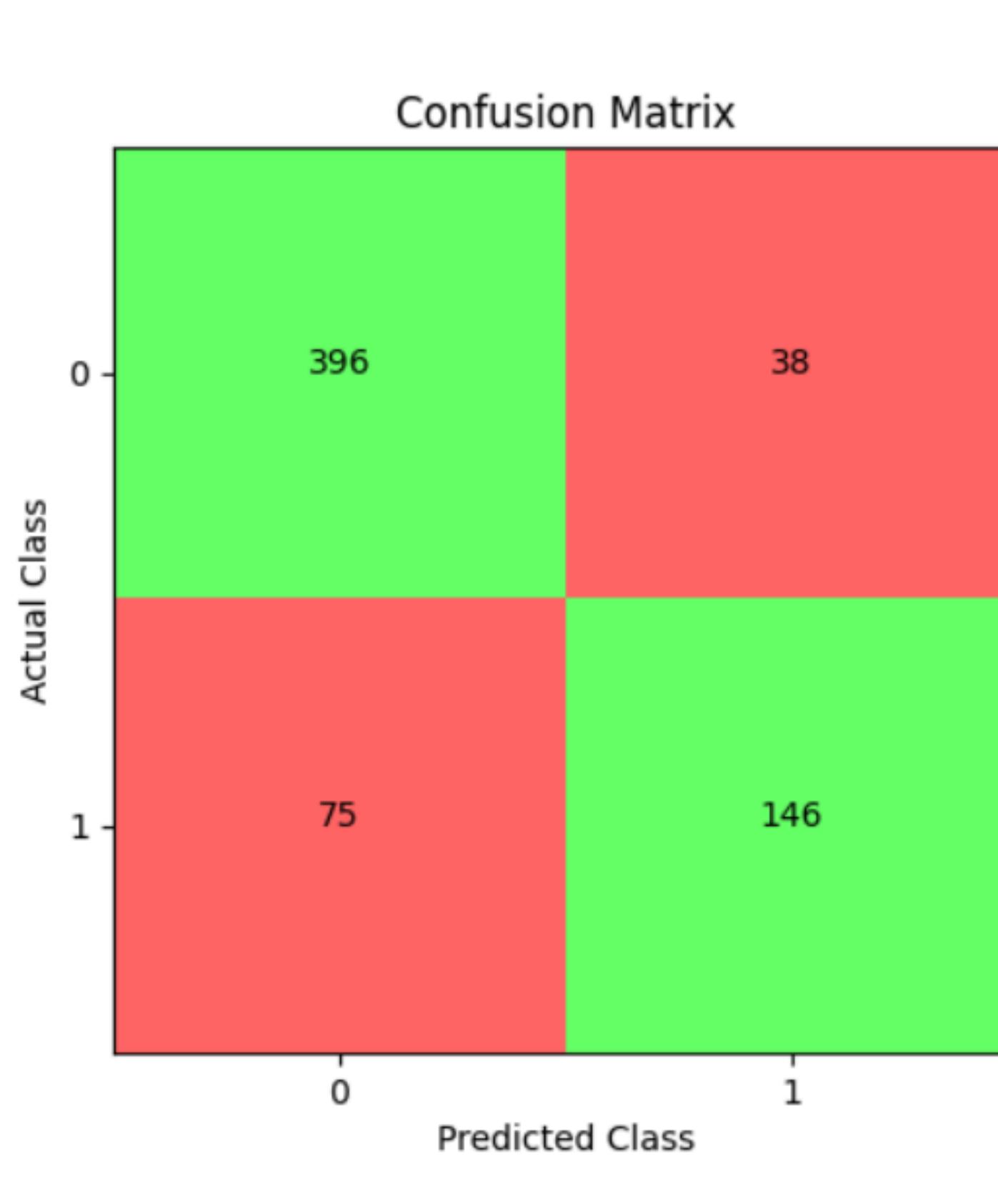
C45



CART



Random forest



Формули показників

- Точність

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

- Чутливість

$$SE=TPR= TP/(TP+FN)$$

- Правильність

$$PR = TP/(TP+FP)$$

- Специфічність

$$SP = TNR = TN/(TN+FP)$$

- ϕ_1 - міра

$$\phi_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

Точність

25

125

4 000

id3

1.0

1.0

0.5770992

C45

1.0

1.0

0.796946564

CART

1.0

1.0

0.8045802

Random forest

1.0

1.0

0.8274809

Інші показники

Чутливість

Правильність

Специфічність

F1-оцінка

id3

0.529411764

0.78

0.923963133

0.6307277628

C45

0.6380090497

0.726804123

0.8778801843

0.6795180722

CART

0.66063348416

0.7336683417

0.87788018433

0.6952380952

Random forest

0.66063348416

0.79347826086

0.91244239631

0.7209876543

Висновки

Час

Точність

Чутливість

Правильність

Специфічність

F1-оцінка

ID3



C45



CART



Random Forest

