**ЛАБОРАТОРНА РОБОТА № 2**

**ПОРІВНЯННЯ МЕТОДІВ КЛАСИФІКАЦІЇ ДАНИХ**

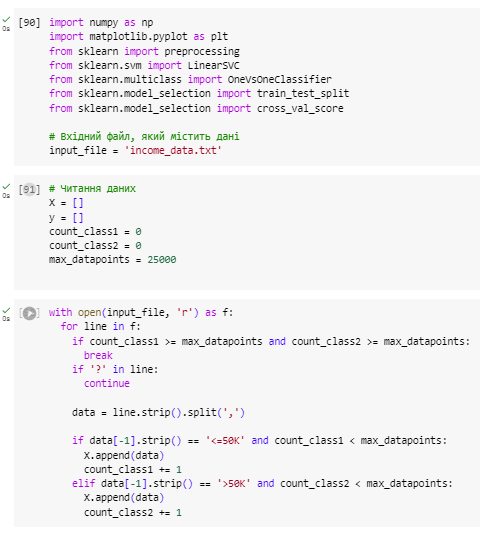
**Мета роботи:** використовуючи спеціалізовані бібліотеки та мову програмування Python дослідити різні методи класифікації даних та навчитися їх порівнювати.

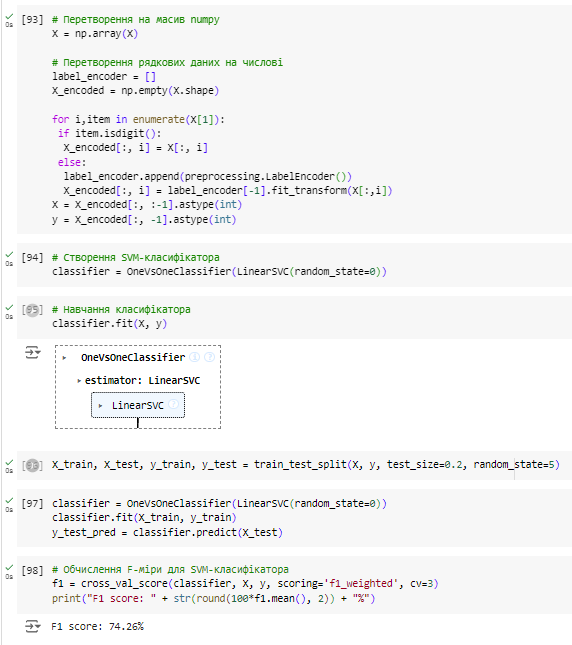
**2. ЗАВДАННЯ НА ЛАБОРАТОРНУ РОБОТУ ТА МЕТОДИЧНІ РЕКОМЕНДАЦІЇ ДО ЙОГО ВИКОНАННЯ**

**Завдання 2.1. Класифікація за допомогою машин опорних векторів (SVM)**

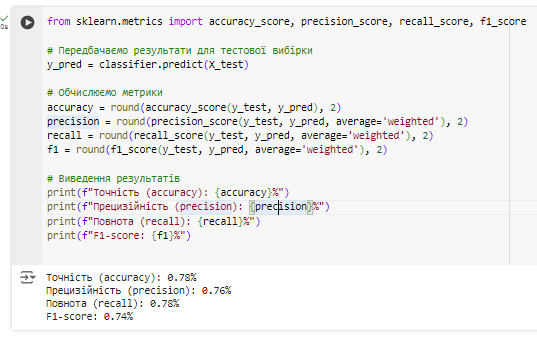
**14 ознак з файлу adult.names - їх назви та що вони позначають та вид.**

1. **age** – числова (continuous).
2. **workclass** – категоріальна: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. **fnlwgt** – числова (continuous).
4. **education** – категоріальна: Bachelors, Some-college, HS-grad, Prof-school, та інші.
5. **education-num** – числова (continuous), кількість років освіти.
6. **marital-status** – категоріальна: Married-civ-spouse, Divorced, Never-married, та іншв.
7. **occupation** – категоріальна: Tech-support, Sales, Exec-managerial, Prof-specialty, та інші.
8. **relationship** – категоріальна: Wife, Own-child, Husband, Not-in-family, та інші.
9. **race** – категоріальна: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. **sex** – категоріальна: Female, Male.
11. **capital-gain** – числова (continuous).
12. **capital-loss** – числова (continuous).
13. **hours-per-week** – числова (continuous).
14. **native-country** – категоріальна: United-States, Cambodia, England, та інші.

**2.1.1. Ознайомтесь з набором даних. **

****

**Обчисліть значення інших показників якості класифікації (акуратність, повнота, точність) та разом з F1 занесіть їх у звіт. (Див. ЛР-1).**

****

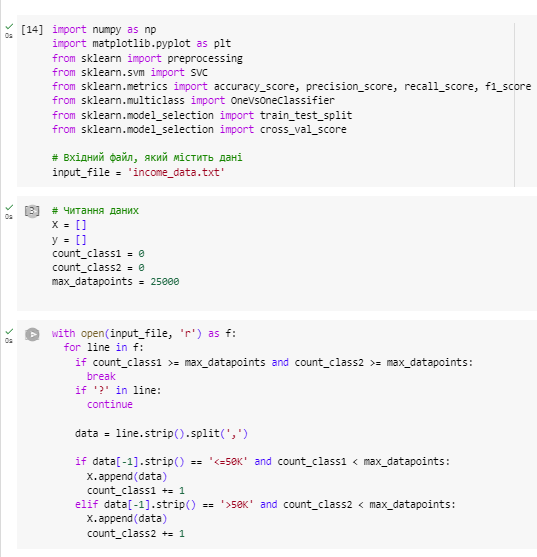
**Висновок:**

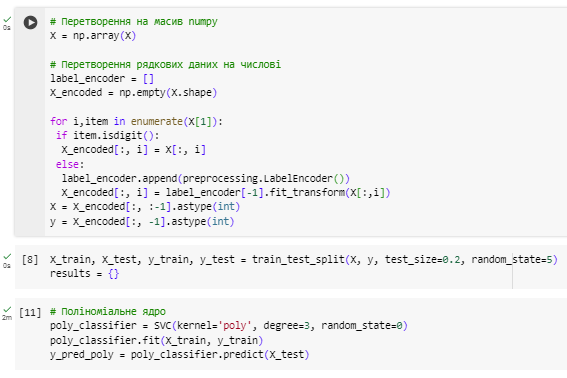
У завданні описується процес створення класифікатора для задачі прогнозування доходу на основі набору даних. Ось основні кроки:

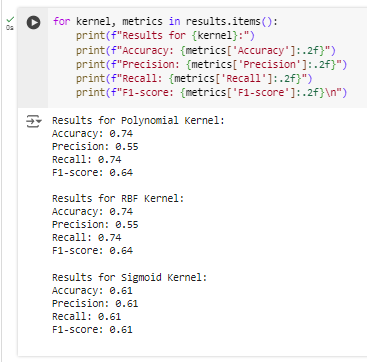
1. **Завантаження даних**: Дані зчитуються з файлу income\_data.txt, після чого відбираються дві класи: <=50K та >50K, кожен з яких обмежується кількістю 25,000 зразків.
2. **Кодування даних**: За допомогою LabelEncoder рядкові дані перетворюються на числові для подальшої обробки.
3. **Побудова класифікатора**: Використовується модель SVM, зокрема OneVsOneClassifier разом з LinearSVC, щоб навчити класифікатор на цих даних.
4. **Прогнозування тестової точки**: Тестова точка також кодується, після чого класифікатор передбачає, що вона належить до класу <=50K.

Отже, тестова точка була віднесена до класу <=50K згідно з результатом класифікатора.

**Завдання 2.2. Порівняння якості класифікаторів SVM з нелінійними ядрами**

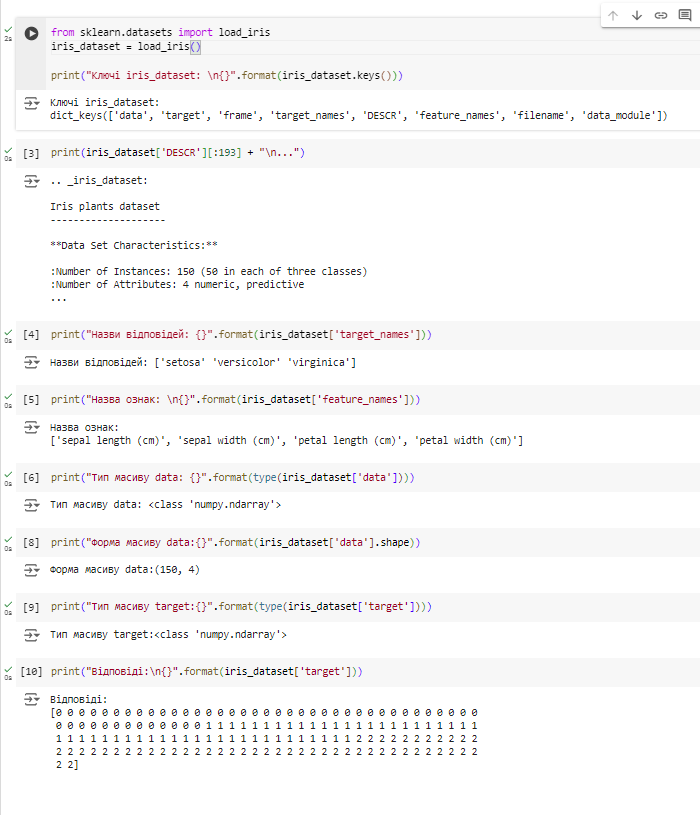
****

****

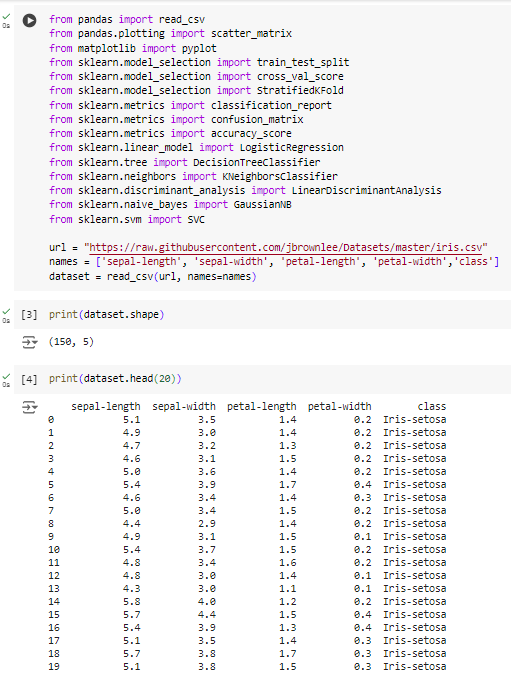
****

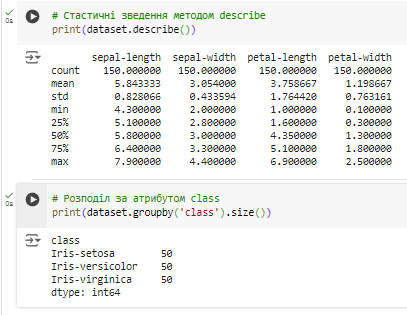
**Завдання 2.3. Порівняння якості класифікаторів на прикладі класифікації сортів ірисів**

КРОК 1. ЗАВАНТАЖЕННЯ ТА ВИВЧЕННЯ ДАНИХ

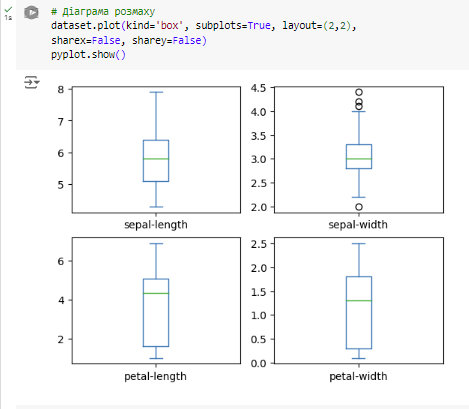
****

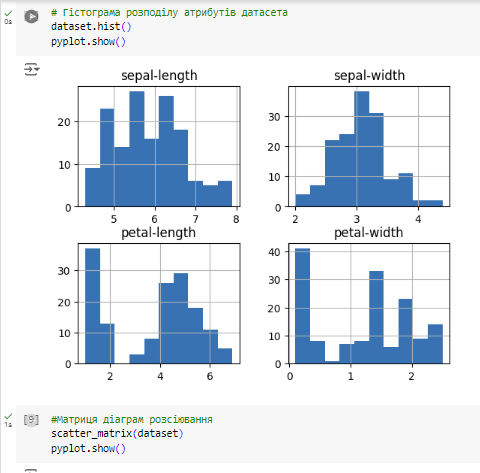
Є ще кілька варіантів завантаження наборів даних: з таблиці або з URL адреси зберігання.

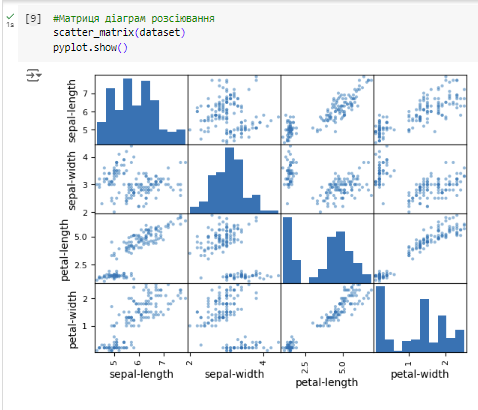




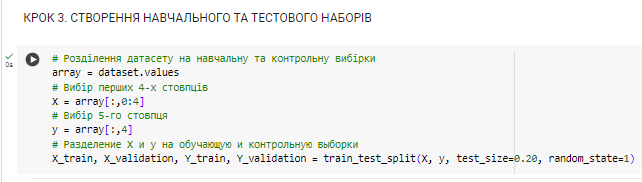
КРОК 2. ВІЗУАЛІЗАЦІЯ ДАНИХ





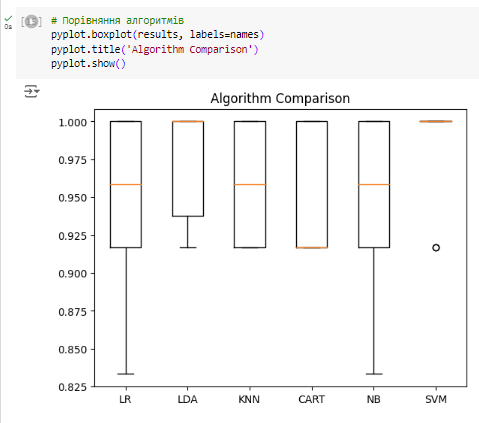


КРОК 3. СТВОРЕННЯ НАВЧАЛЬНОГО ТА ТЕСТОВОГО НАБОРІВ

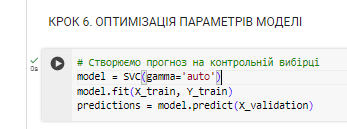


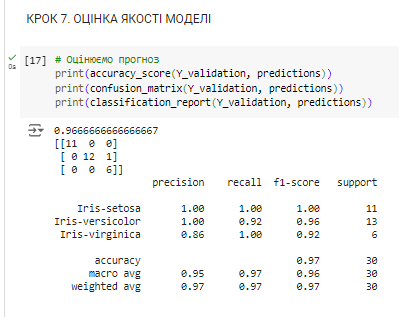
КРОК 4. КЛАСИФІКАЦІЯ (ПОБУДОВА МОДЕЛІ)

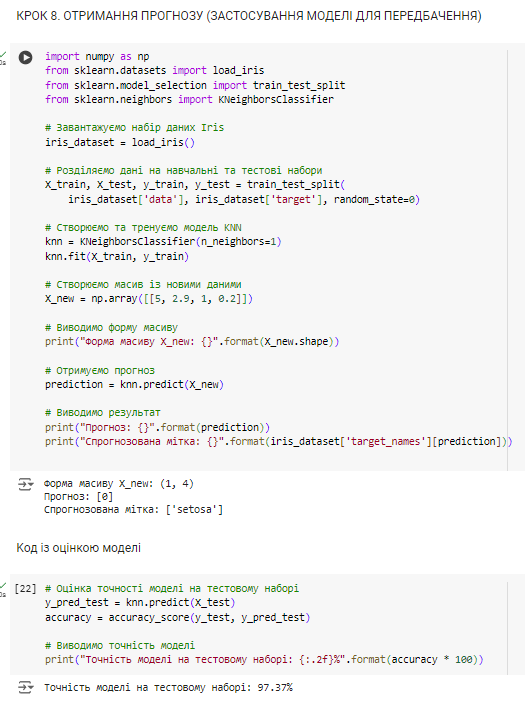




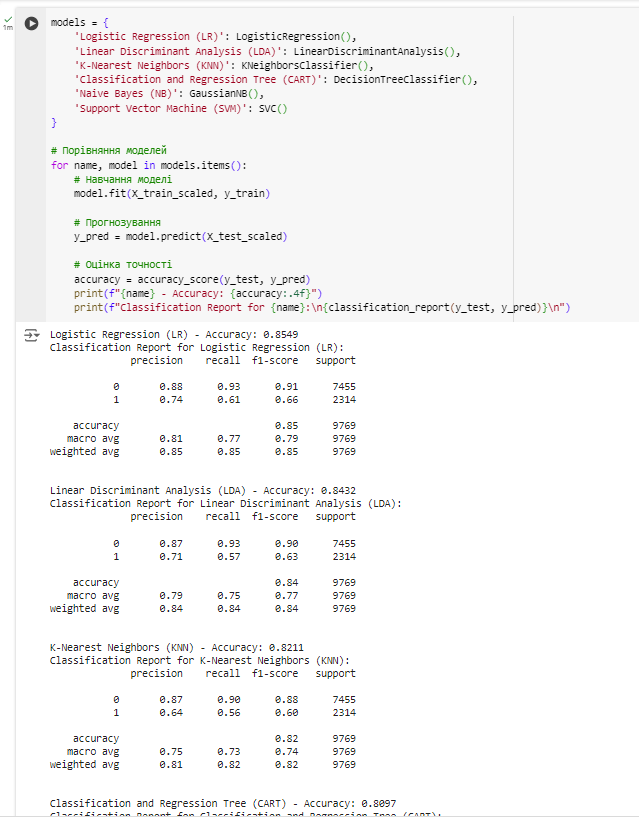
КРОК 6. ОТРИМАННЯ ПРОГНОЗУ (ПЕРЕДБАЧЕННЯ НА ТРЕНУВАЛЬНОМУ НАБОРІ)

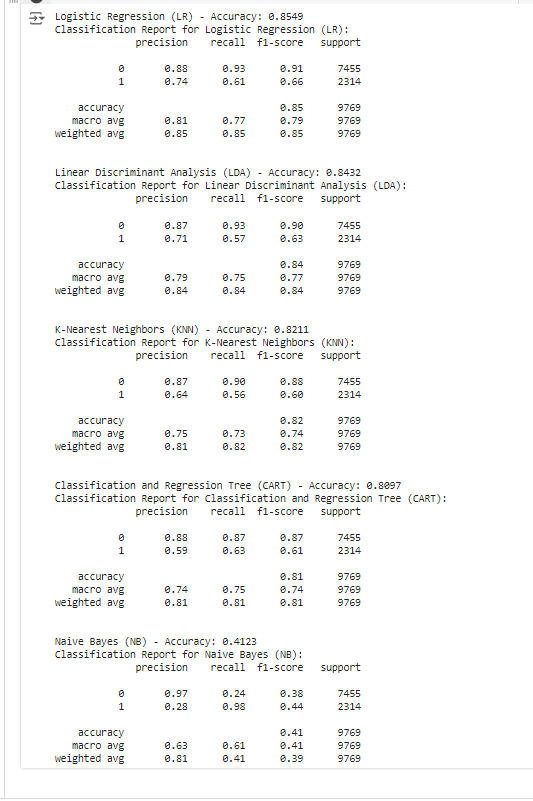


КРОК 7. ОЦІНКА ЯКОСТІ МОДЕЛІ

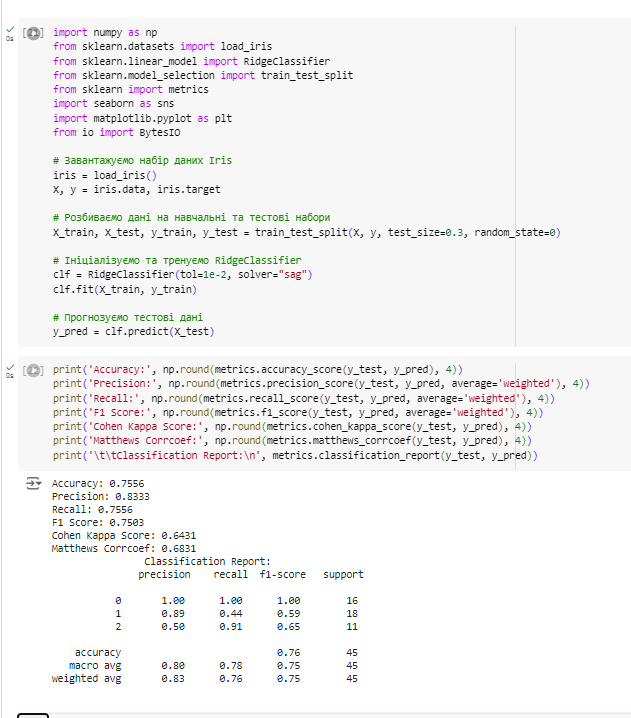
КРОК 8. ОТРИМАННЯ ПРОГНОЗУ (ЗАСТОСУВАННЯ МОДЕЛІ ДЛЯ ПЕРЕДБАЧЕННЯ) 

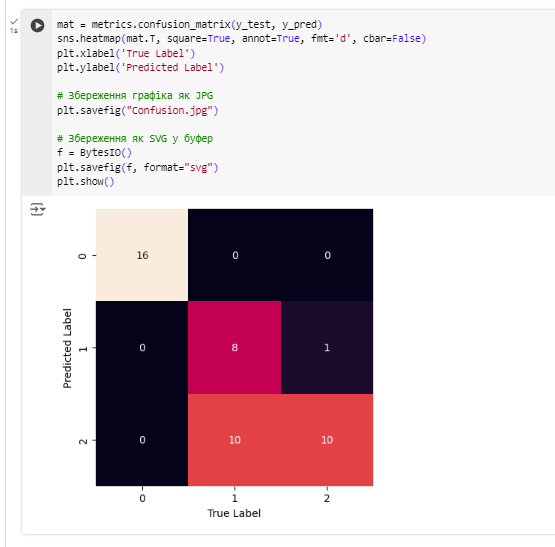
Завдання 2.4. Порівняння якості класифікаторів для набору даних завдання 2.1





Завдання 2.5. Класифікація даних лінійним класифікатором Ridge





**Пояснення налаштувань класифікатора Ridge**

1. **tol=1e-2** — це параметр точності, який визначає мінімальне значення зміни коефіцієнтів моделі під час оптимізації. Чим менше це значення, тим точніше проводиться оптимізація, але це може збільшити час роботи.
2. **solver='sag'** — це метод оптимізації, що використовується для підгонки моделі. SAG (Stochastic Average Gradient) — це швидкий метод для лінійних моделей, який добре працює для великих наборів даних.

**Пояснення показників якості**

1. **Accuracy (Точність)**: Загальна частка правильно передбачених класів серед усіх тестових прикладів. У нашому випадку це показує загальну якість моделі для трьох класів.
2. **Precision (Точність)**: Відсоток правильних позитивних передбачень з усіх передбачених позитивів. Враховує, скільки разів модель правильно передбачила певний клас.
3. **Recall (Повнота)**: Відсоток правильних позитивних передбачень з усіх реальних позитивів. Враховує, наскільки добре модель розпізнає позитивні випадки.
4. **F1 Score**: Гармонійне середнє між точністю та повнотою. Важливий показник, коли між точністю і повнотою є компроміс.
5. **Cohen Kappa Score**: Коефіцієнт Коена Каппа вимірює рівень угоди між двома класифікаторами, враховуючи випадкові збіги. Він дає показник від -1 (повна незгода) до 1 (повна згода).
6. **Matthews Corrcoef**: Коефіцієнт кореляції Метьюза (MCC) — це метрика, яка враховує всі чотири значення з матриці плутанини (TP, TN, FP, FN) і дає оцінку від -1 до 1. Це один з найточніших способів оцінки якості класифікатора, особливо при дисбалансі класів.

**На зображенні Confusion.jpg показано матрицю плутанини (confusion matrix)**, яка використовується для оцінки ефективності класифікаційної моделі. Вона показує, як добре модель класифікує зразки у відповідні класи, порівнюючи передбачені значення з фактичними (істинними) значеннями.

**Матриця плутанини**

Матриця плутанини (confusion matrix) показує кількість правильних і неправильних передбачень для кожного класу. Кожен рядок представляє реальні значення, а кожен стовпець — передбачені значення. На графіку ми бачимо:

* Клітини на діагоналі представляють кількість правильно передбачених значень для кожного класу.
* Всі інші клітини — це неправильні передбачення (помилки моделі).

Зображення **Confusion.jpg** відображає точність моделі у вигляді матриці плутанини для кожного класу (три класи Iris).

**Пояснення коефіцієнтів Коена Каппа та Метьюза**

1. **Коефіцієнт Коена Каппа** вимірює угоду між двома класифікаторами, враховуючи можливість випадкових угод. Значення близьке до 1 означає високу узгодженість, а значення близьке до 0 означає, що угода не краща за випадкову.
2. **Коефіцієнт кореляції Метьюза** (MCC) — це метрика, яка враховує всі можливі типи передбачень (TP, TN, FP, FN). Вона особливо корисна при нерівномірному розподілі класів, оскільки дає збалансовану оцінку якості класифікації. Значення MCC варіюються від -1 (повна невідповідність) до 1 (ідеальна відповідність).