**ЛАБОРАТОРНА РОБОТА № 1**

**ПОПЕРЕДНЯ ОБРОБКА ТА КОНТРОЛЬОВАНА**

**КЛАСИФІКАЦІЯ ДАНИХ**

***Мета роботи:*** використовуючи спеціалізовані бібліотеки та мову програмування Python дослідити попередню обробку та класифікацію даних.

**Завдання 2.1.1 Бінарізація.**

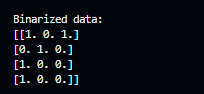


Рис.1 Результат виконання завдання 2.1.1

**Завдання 2.1.2 Виключення середнього.**

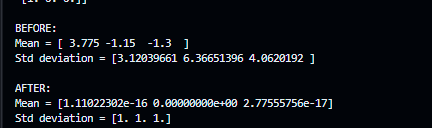


Рис.2 Результат виконання завдання 2.1.2

**Завдання 2.1.3 Масштабування.**

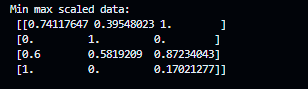


Рис.3 Результат виконання завдання 2.1.3

**Завдання 2.1.4 Нормалізація.**

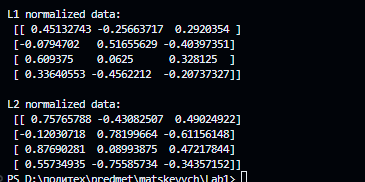
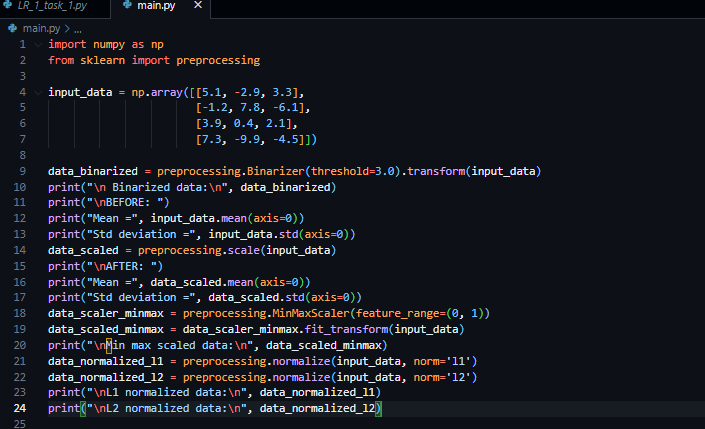


Рис.4 Результат виконання завдання 2.1.4



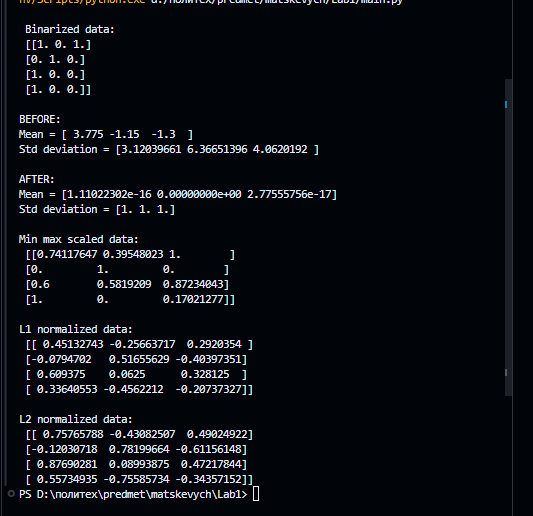


Рис. 5 - 6 Результат виконання завдання 2.1.1 – 2.1.4

Як бачимо, L1-нормалізація менш чутлива до викидів.

**Завдання 2.1.5 Кодування міток.**

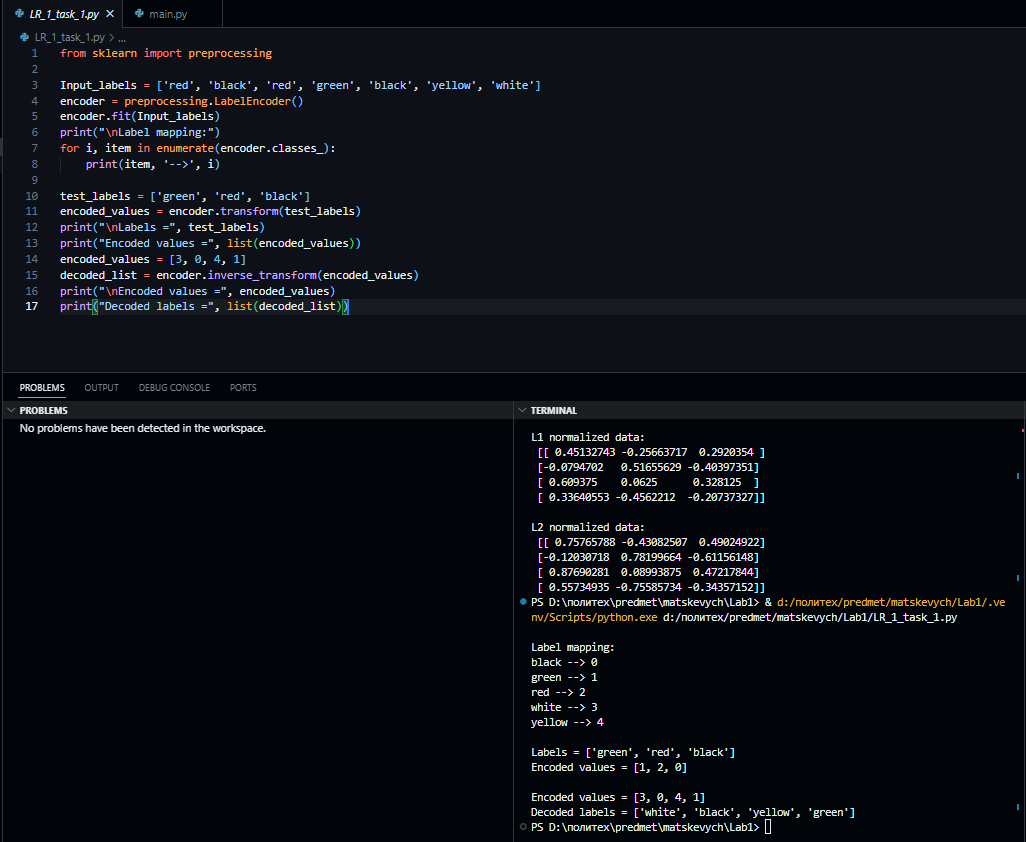


Рис. 7. Результат виконання завдання 2.1.5

**Завдання 2.2 Попередня обробка нових даних.**



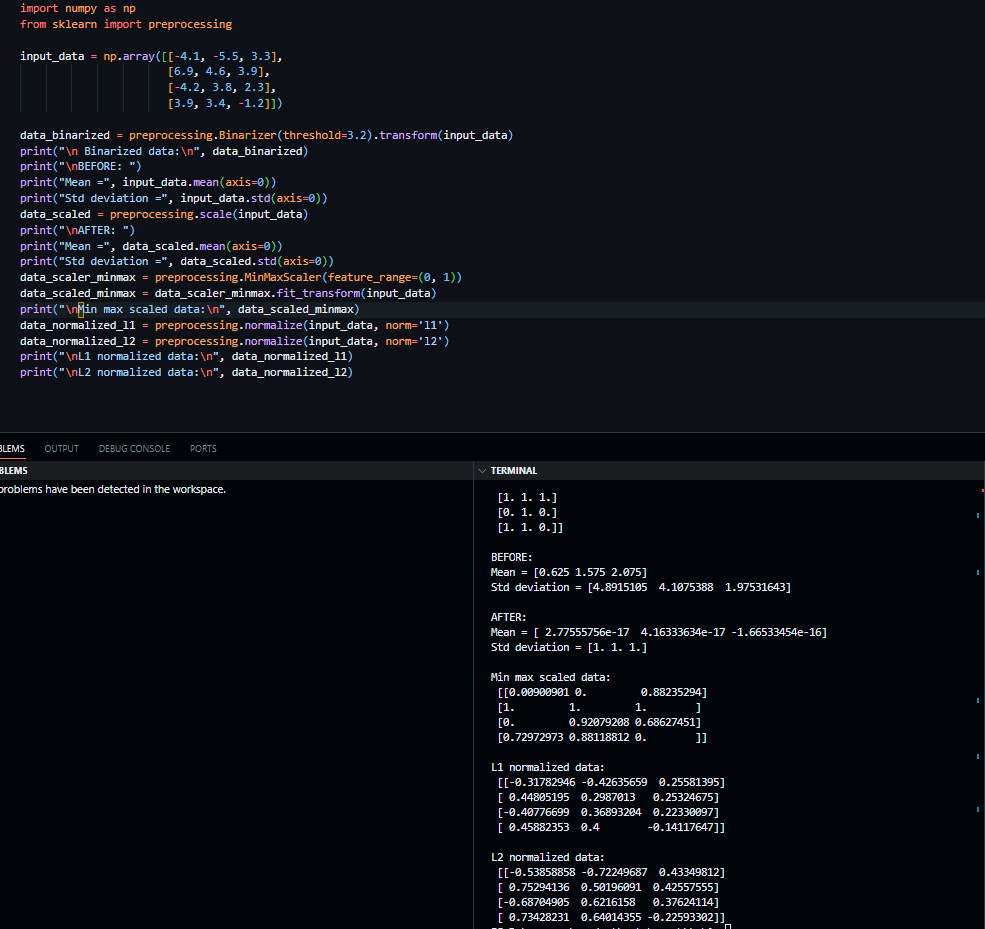


Рис.8. Результат виконання завдання 2.2

**Завдання 2.3 Класифікація логістичною регресією або логістичний**

**класифікатор.**

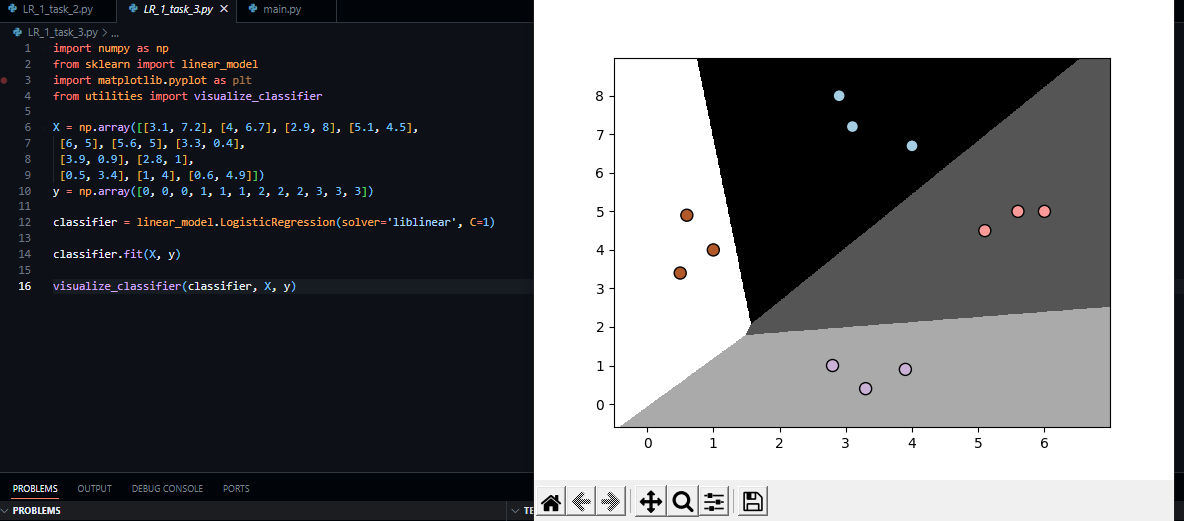
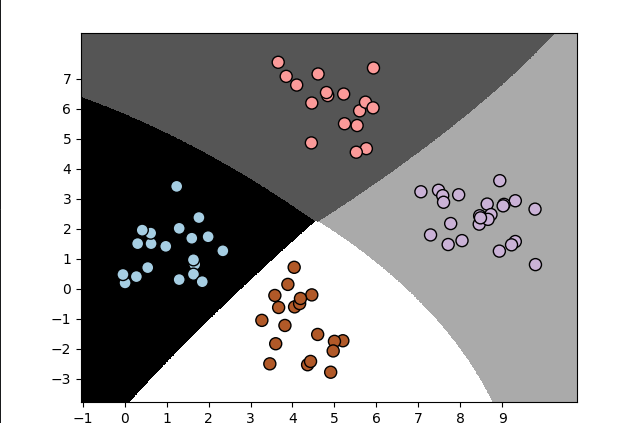
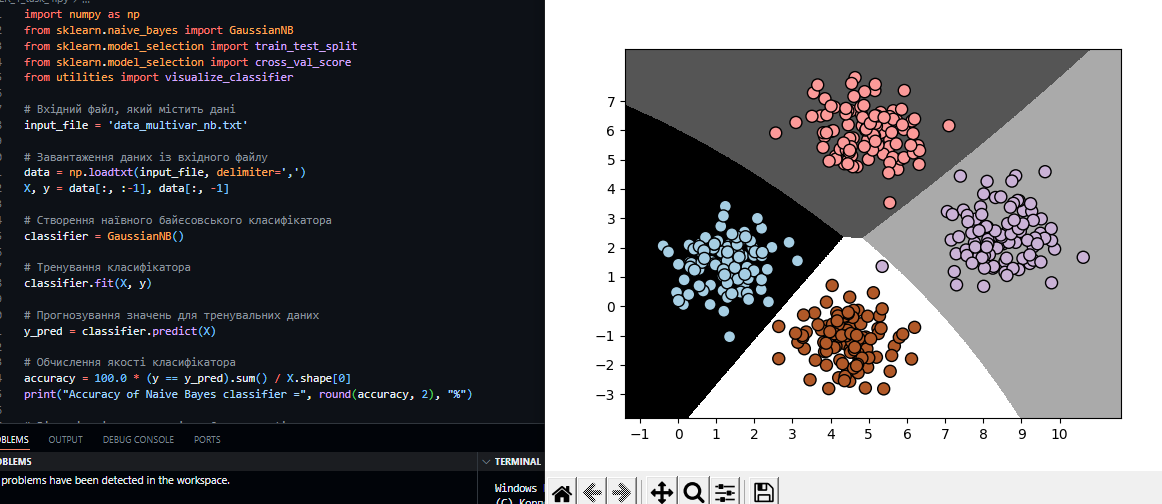


Рис.9 Результат виконання завдання 2.3

**Завдання 2.4 Класифікація наївним байєсовським класифікатором.**





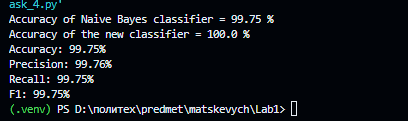


Рис. 10 - 12. Результат виконання завдання 2.4

**Завдання 2.5 Вивчити метрики якості класифікації.**

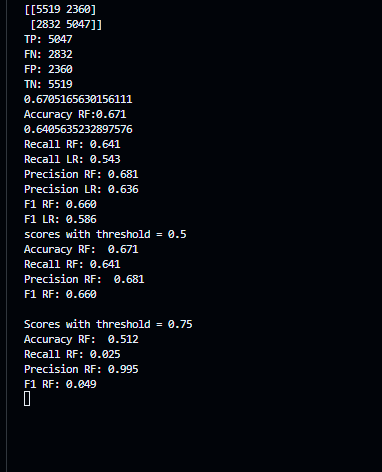


Рис.13 - 15 Результат виконання завдання 2.5

З результатів бачимо, що при збільшенні порогу F1 міра зменшується.

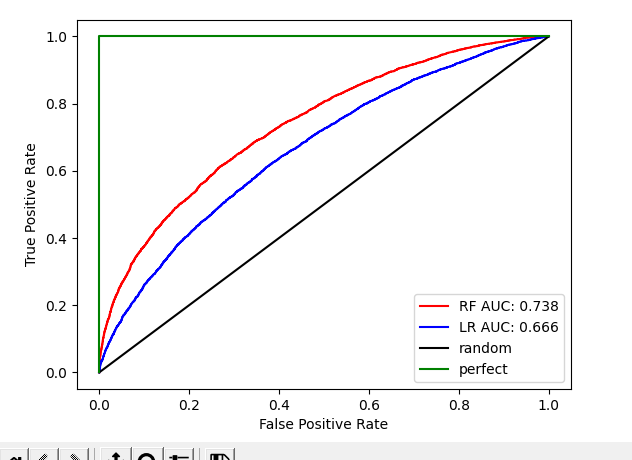


Рис.16 ROC - крива

На графіку бачимо, що RF модель виглядає краще, ніж LR модель , площа під кривою для моделі RF (AUC = 0,738) краще, ніж LR (AUC = 0,666). Проте ефективність кожного з методів залежить від конкретної моделі.

**Завдання 2.6 Розробіть програму класифікації даних в файлі data\_multivar\_nb.txt за допомогою машини опорних векторів (Support Vector Machine - SVМ). Розрахуйте показники якості класифікації. Порівняйте їх з показниками наївного байєсівського класифікатора. Зробіть висновки яку модель класифікації краще обрати і чому.**

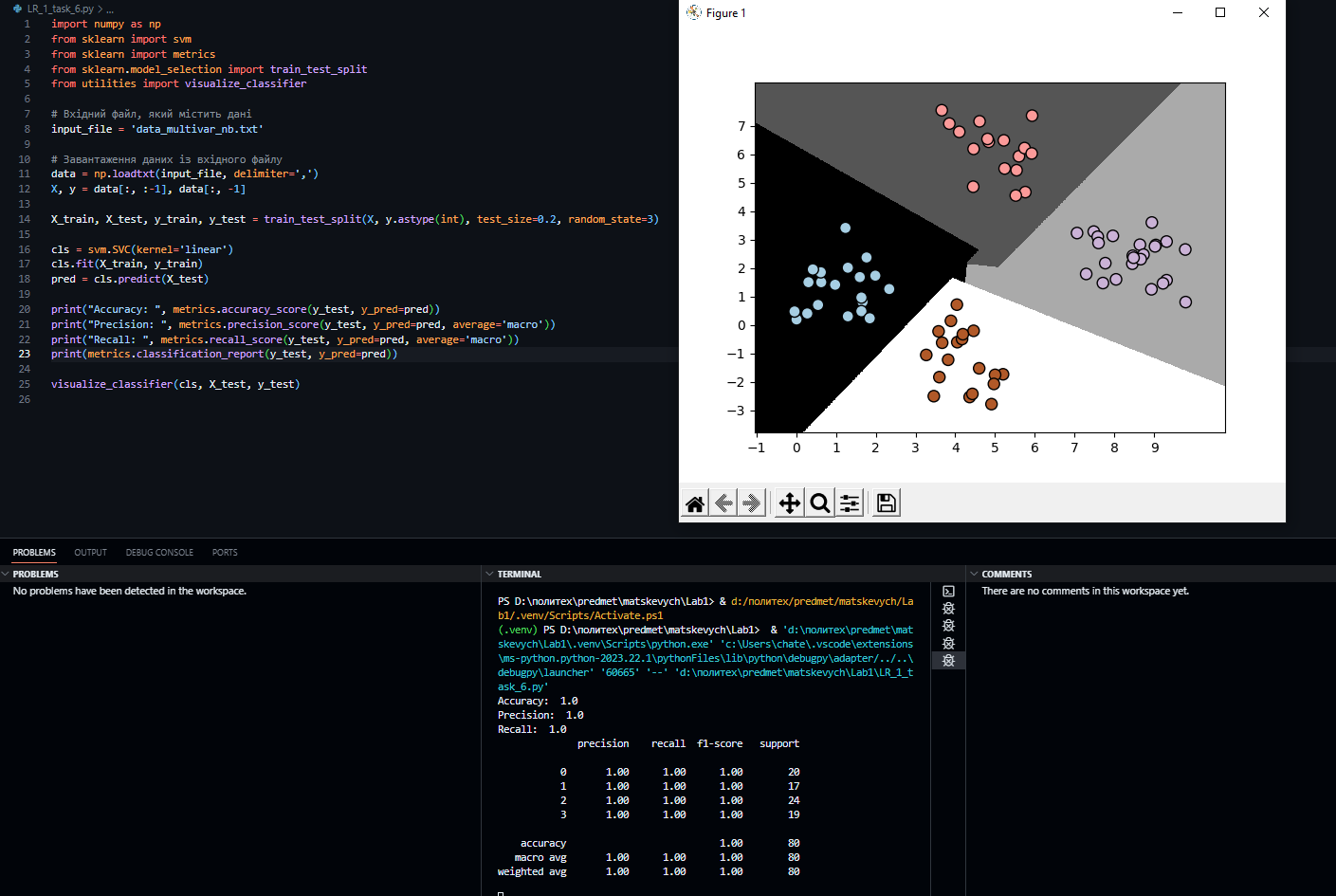


Рис.17 - 18 Результат виконання завдання 2.6

Наївний баєсівський класифікатор і метод опорних векторів (SVM) мають різні параметри, включаючи вибір функції ядра для кожного з них. Обидва алгоритми є дуже чутливими до оптимізації параметрів, тому вибір різних параметрів може суттєво вплинути на отримані результати. Для обраних параметрів NBC працює краще, ніж SVM, що видно в результатах. Проте, за інших параметрів, SVM може показати більш ефективні результати.

**Репозиторій:** https://github.com/RomanMatskevich/SHI.git

**Висновок:** в ході виконання лабораторної роботи використовуючи спеціалізовані бібліотеки та мову програмування Python було досліджено попередню обробку та класифікацію даних.