**ЛАБОРАТОРНА РОБОТА № 5**

**Мета роботи**: використовуючи спеціалізовані бібліотеки та мову програмування Python дослідити методи ансамблів у машинному навчанні.

**Завдання 1.** Створення класифікаторів на основі випадкових та гранично випадкових лісів

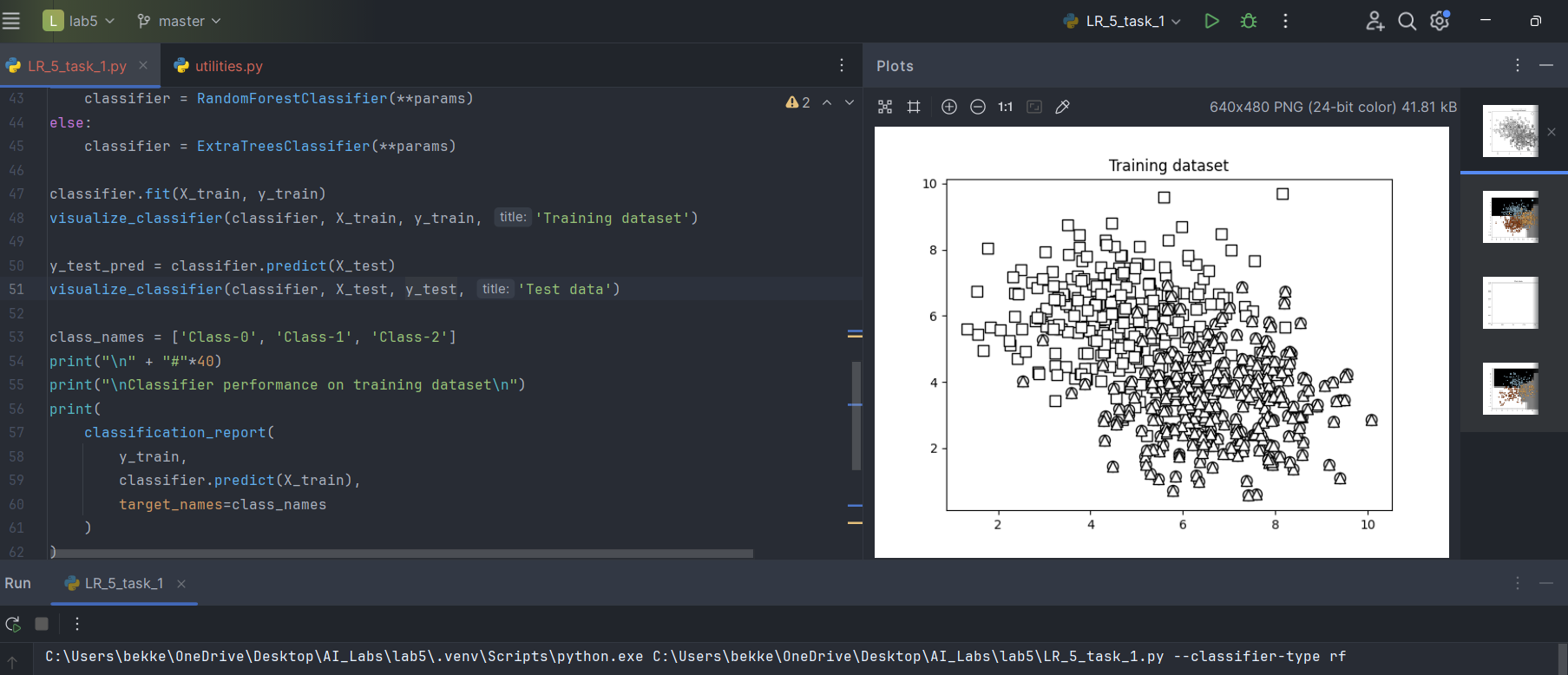
****

Рис 1. Візуалізація вхідних даних

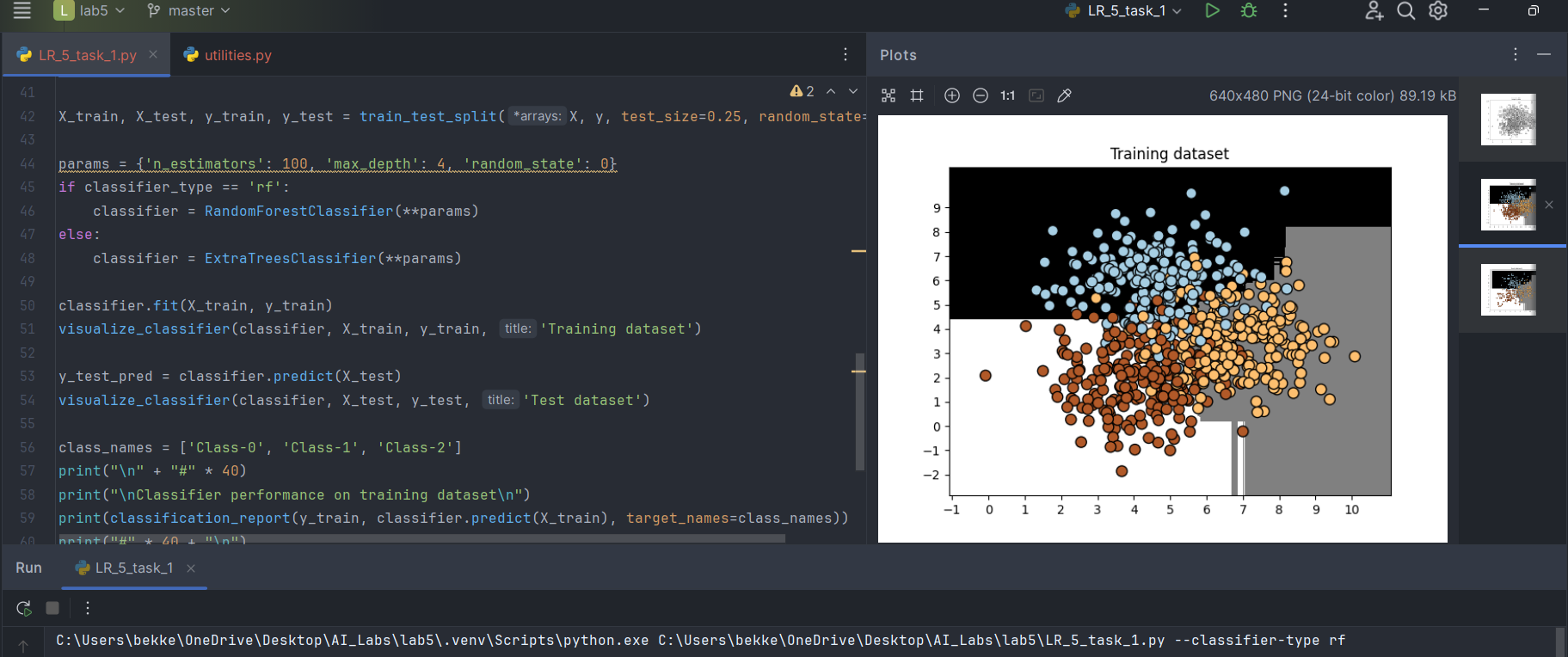
****

Рис 2. Візуалізація класифікації тренувальних даних (rf)

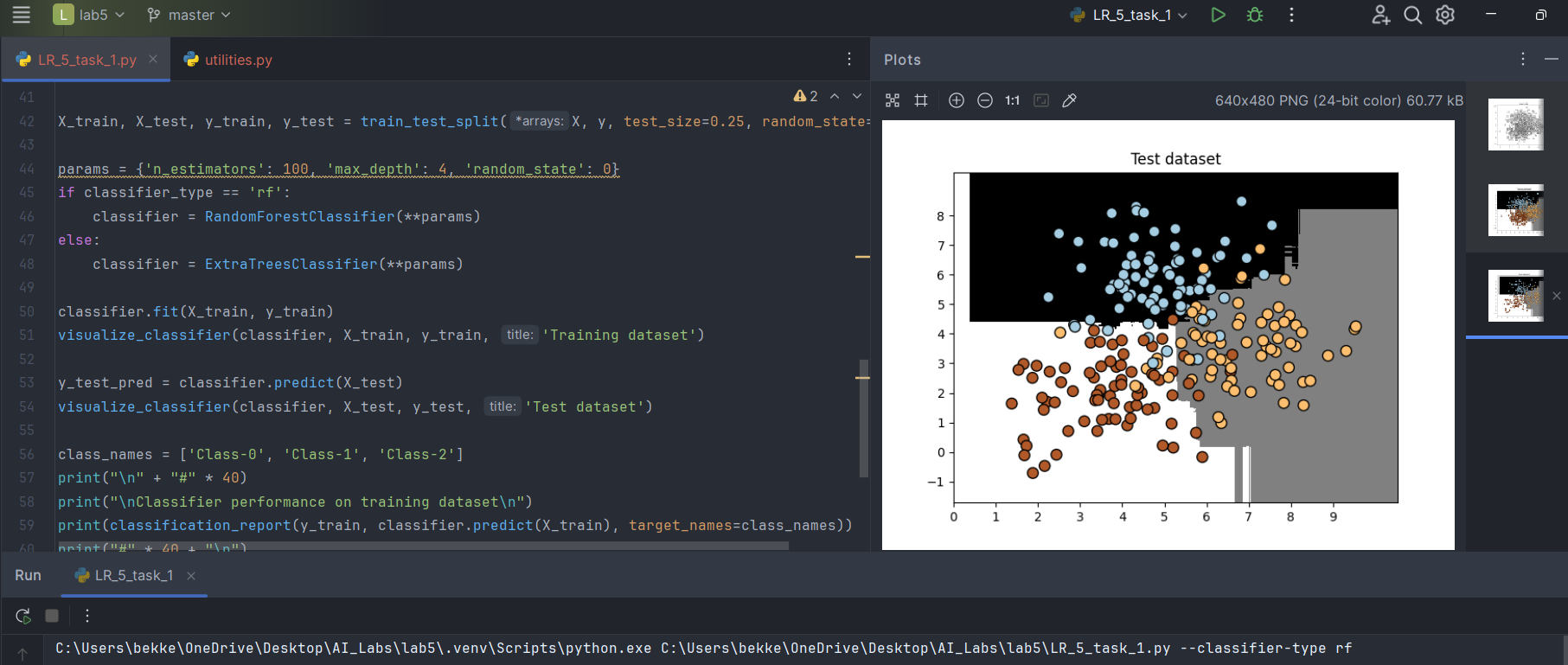


Рис 3. Візуалізація класифікації тестових даних (rf)

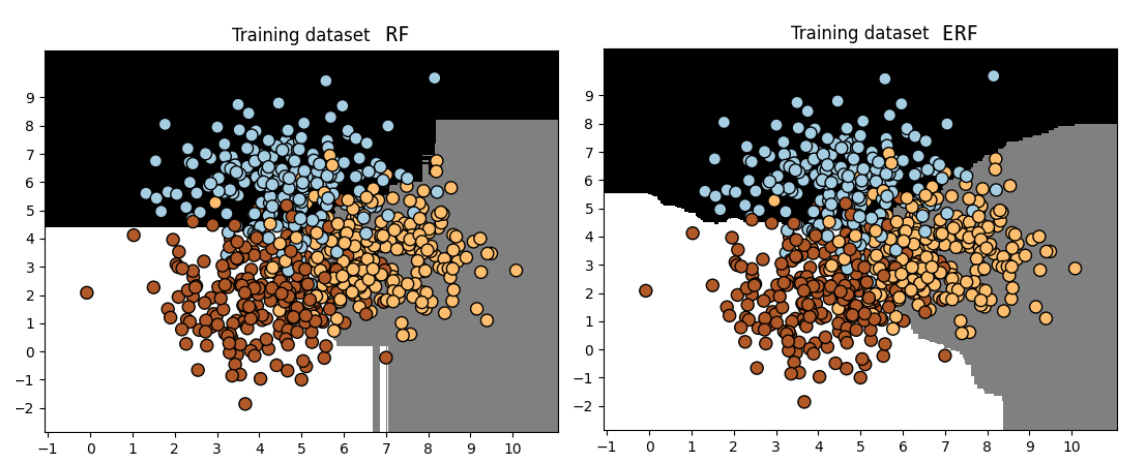
****

Рис 4. Порівняння графіків

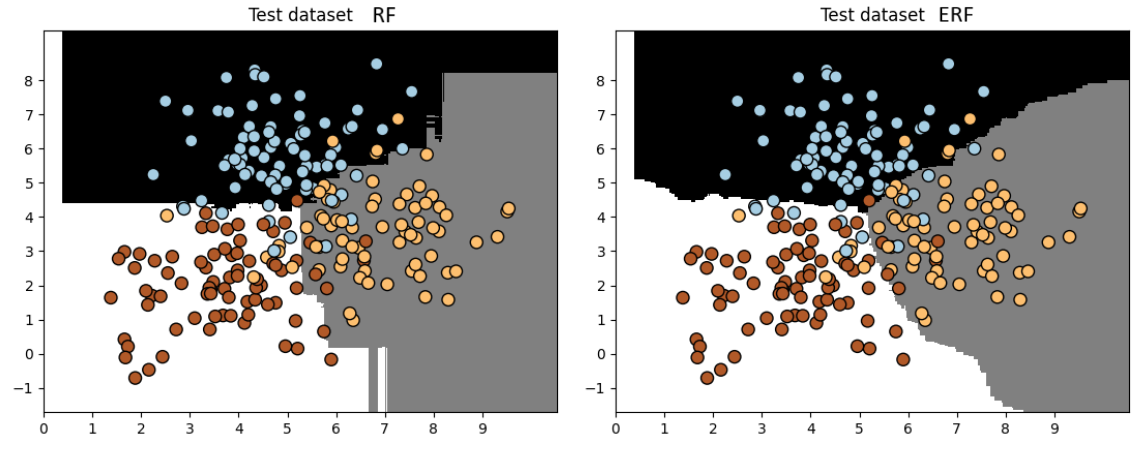


Рис 5. Порівняння графіків

**Оцінка мір достовірності прогнозів**

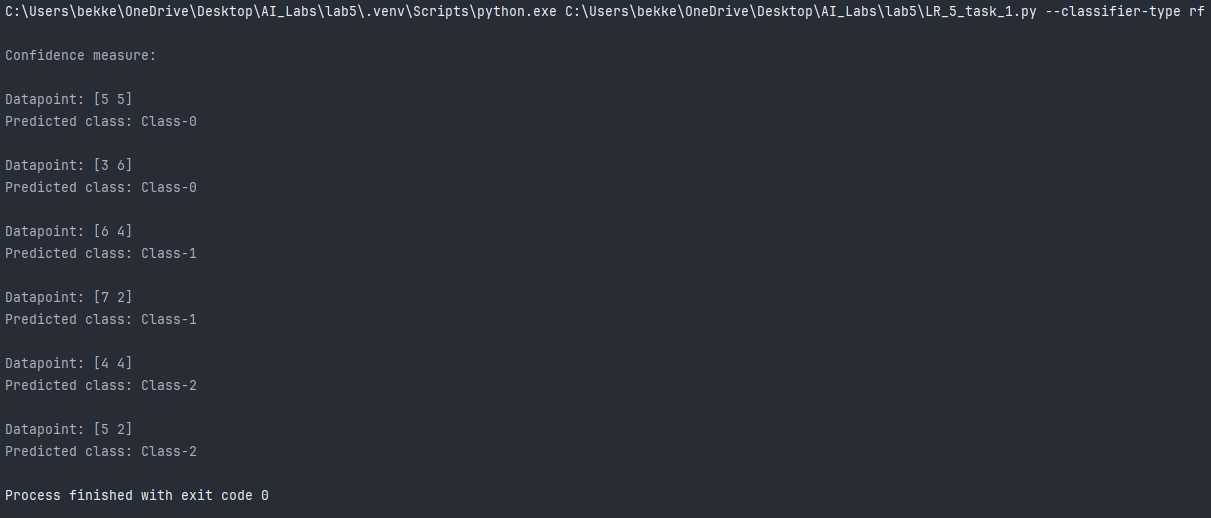
****

Рис 6. Рівні довірливості (rf)

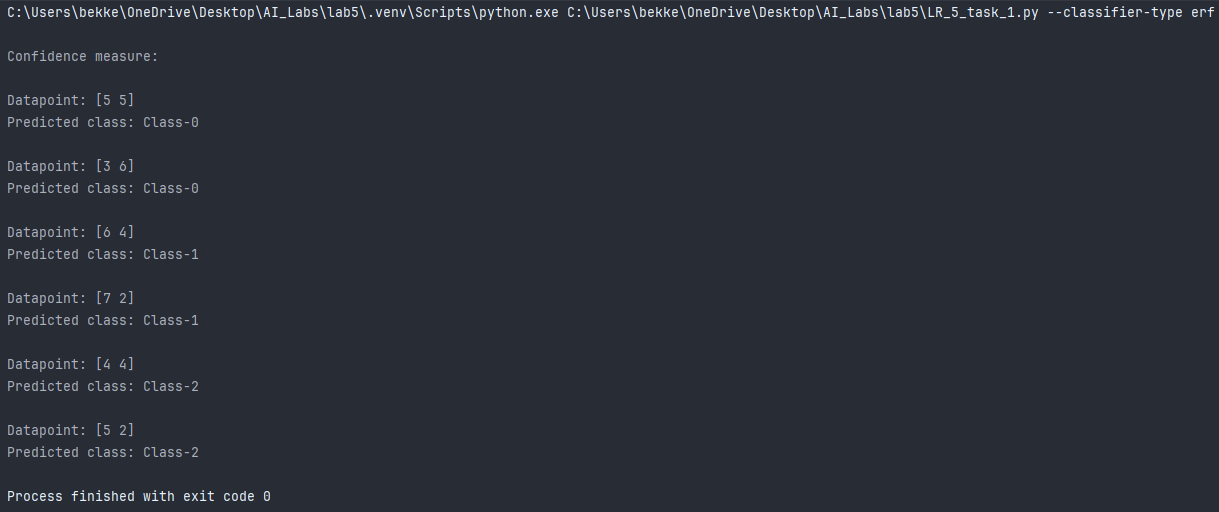


Рис 7. Рівні довірливості (erf)

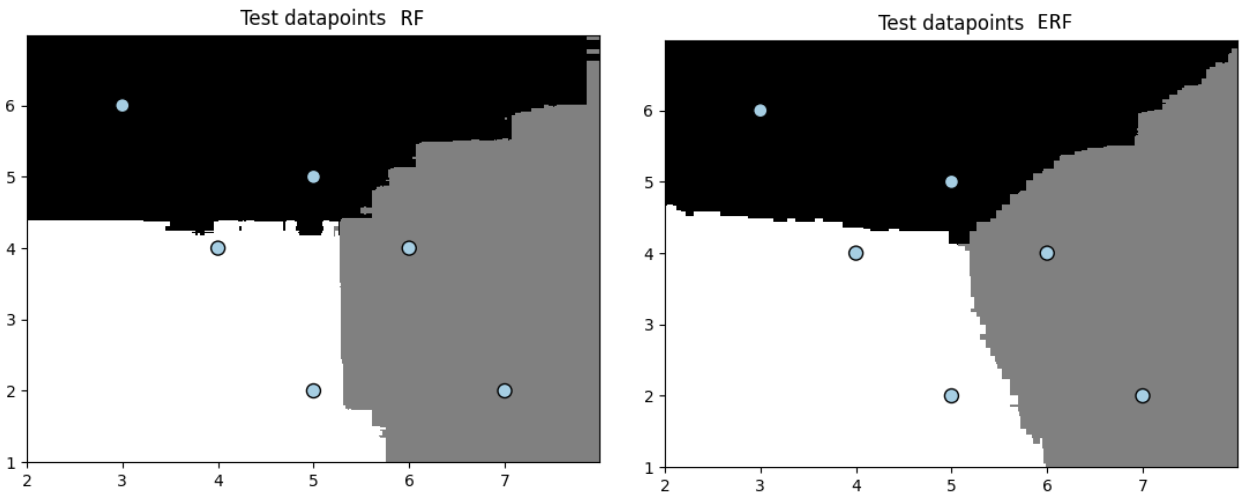


Рис 8. Графіки функцій

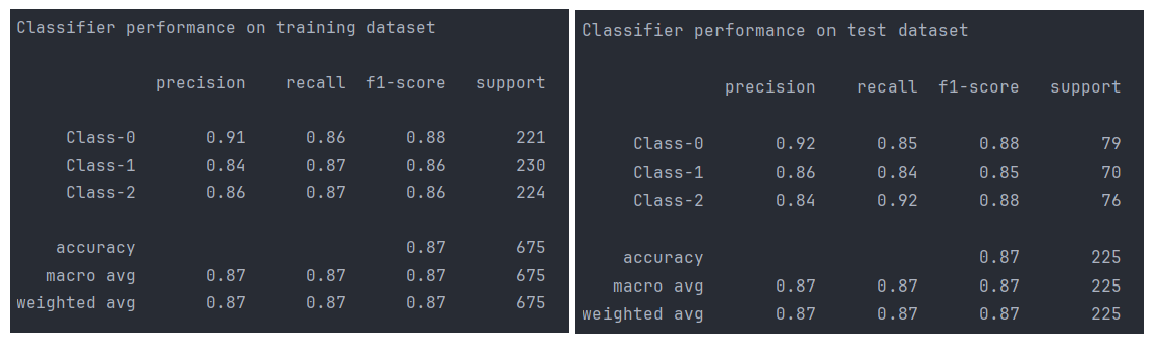


Рис 9. Оцінка якості (RF)

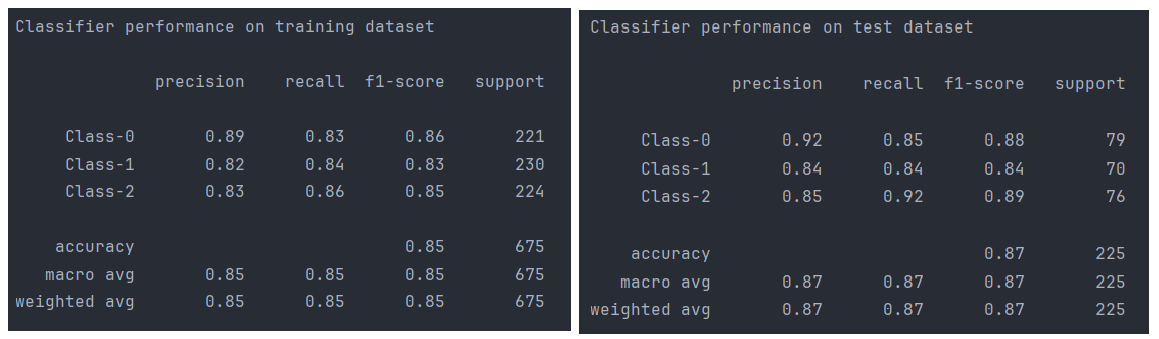


Рис 10. Оцінка якості (ERF)

**Висновок:** на даному етапівдалося порівняти два окремих випадка ансамблевого навчання: «Випадковий ліс» та «Гранично випадковий ліс», за допомогою візуалізації було отримано результати, на яких видно, що дійсно, «Гранично випадковий ліс» призводить до більш гладких меж прийняття рішень, що означає меншу варіативність моделі. Також було отримано результати оцінки якості, що виявились близьким за значенням для обох випадків ансамблевого навчання.

**Завдання 2.** Обробка дисбалансу класів

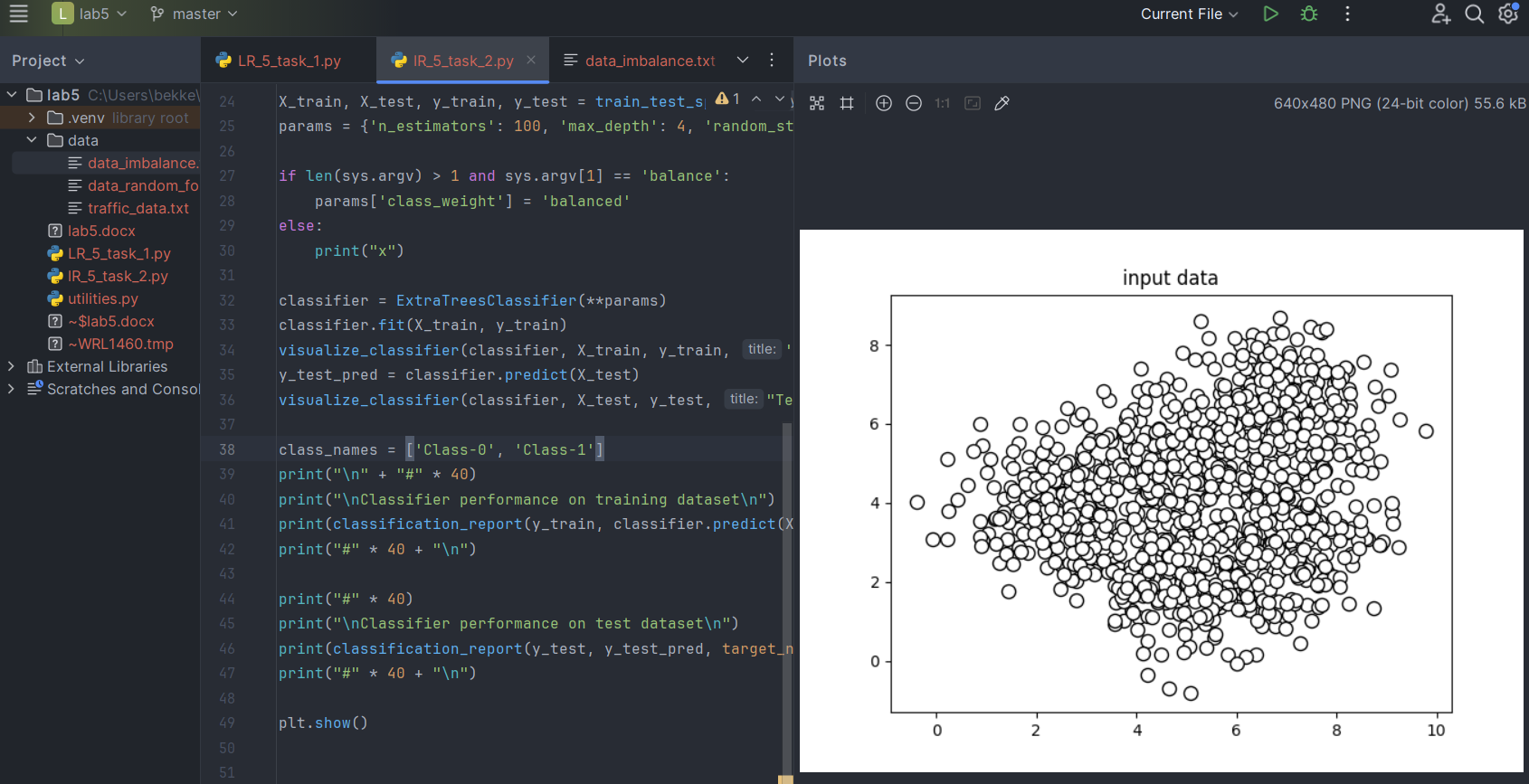
****

Рис 11. Вхідні дані

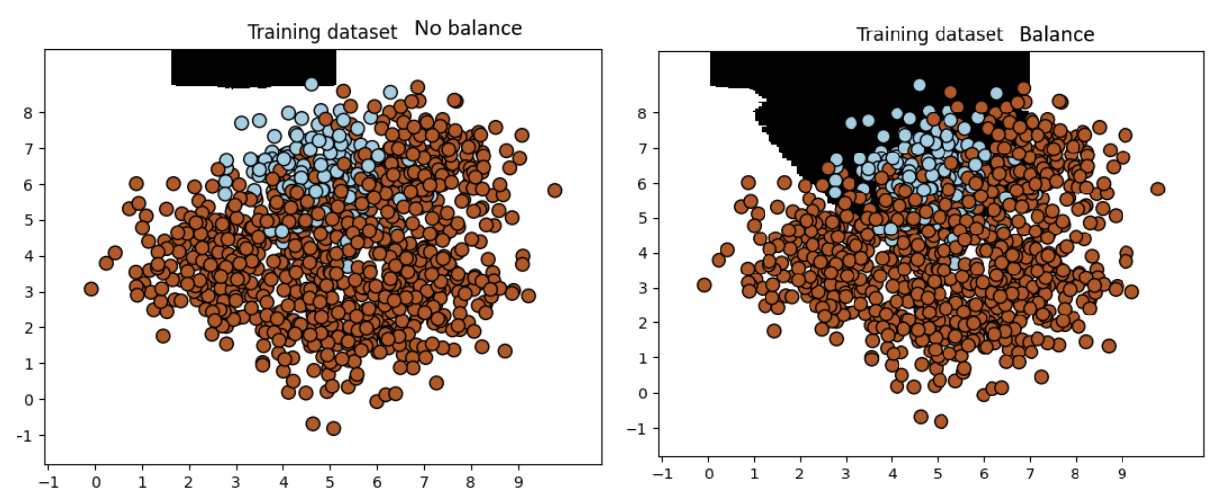
****

Рис 12. Графік даних класифікатора для тестового набору

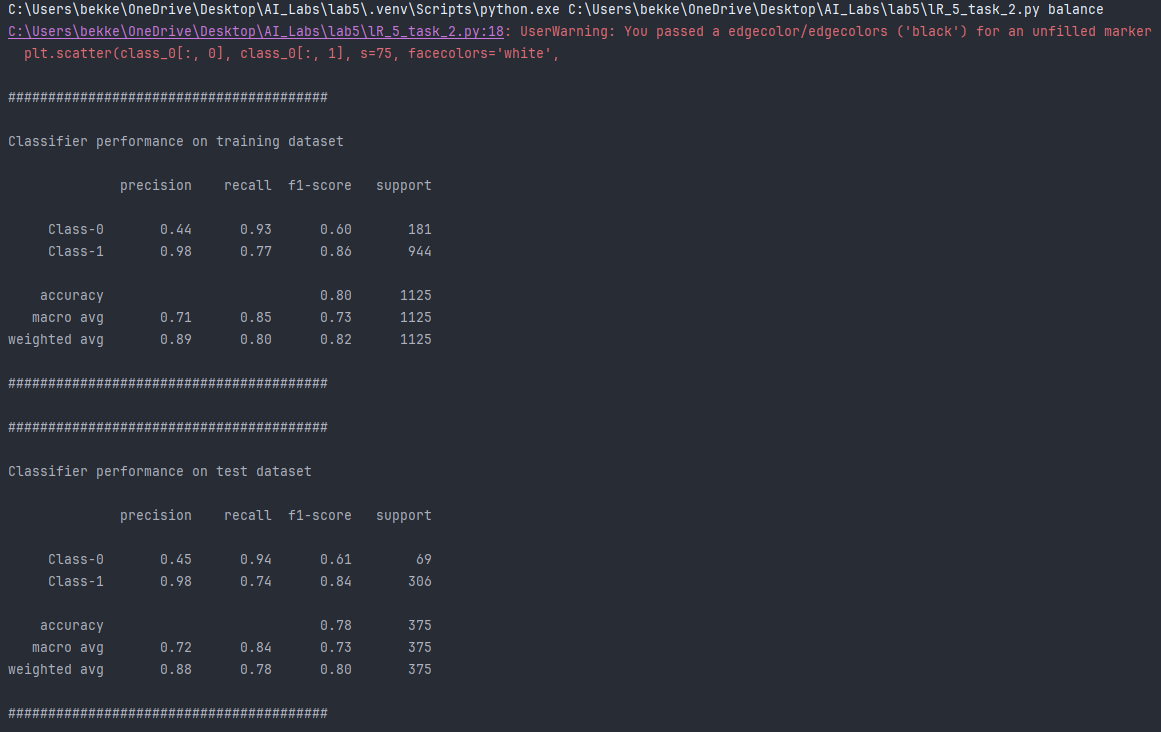


Рис 13. Оцінка якості (with balance)

**Висновок:** за допомогою візуалізації та оцінки якості було доведено, що незбалансовані дані погано впливають на якість роботи класифікатора. Для незбалансованих даних також не вдалося визначити фактичну межу між двома класами (див: рис 13).

**Завдання 3.** Знаходження оптимальних навчальних параметрів за допомогою сіткового пошуку

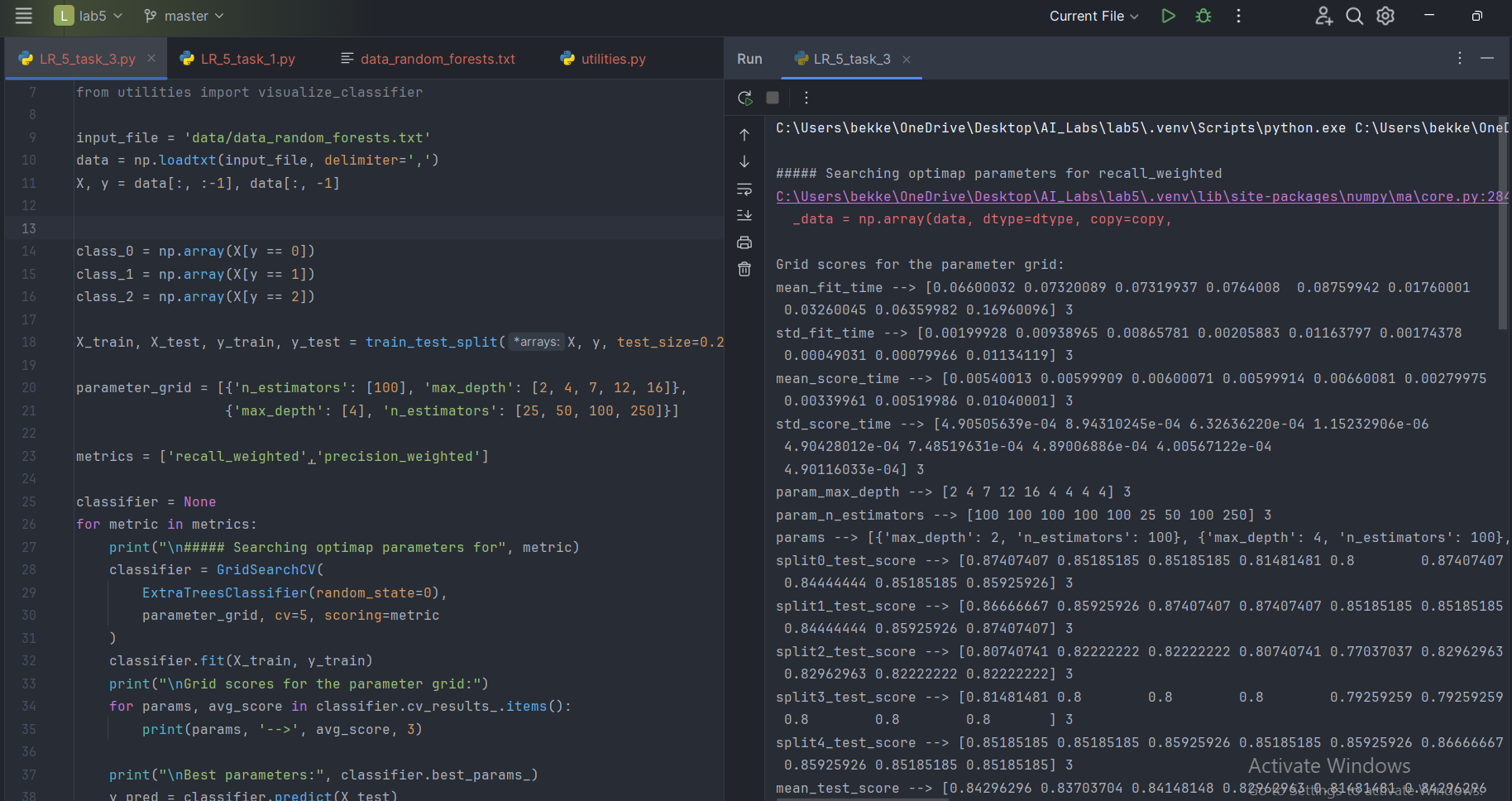


Рис 14. Результати пошуку оптимальних параметрів для recall

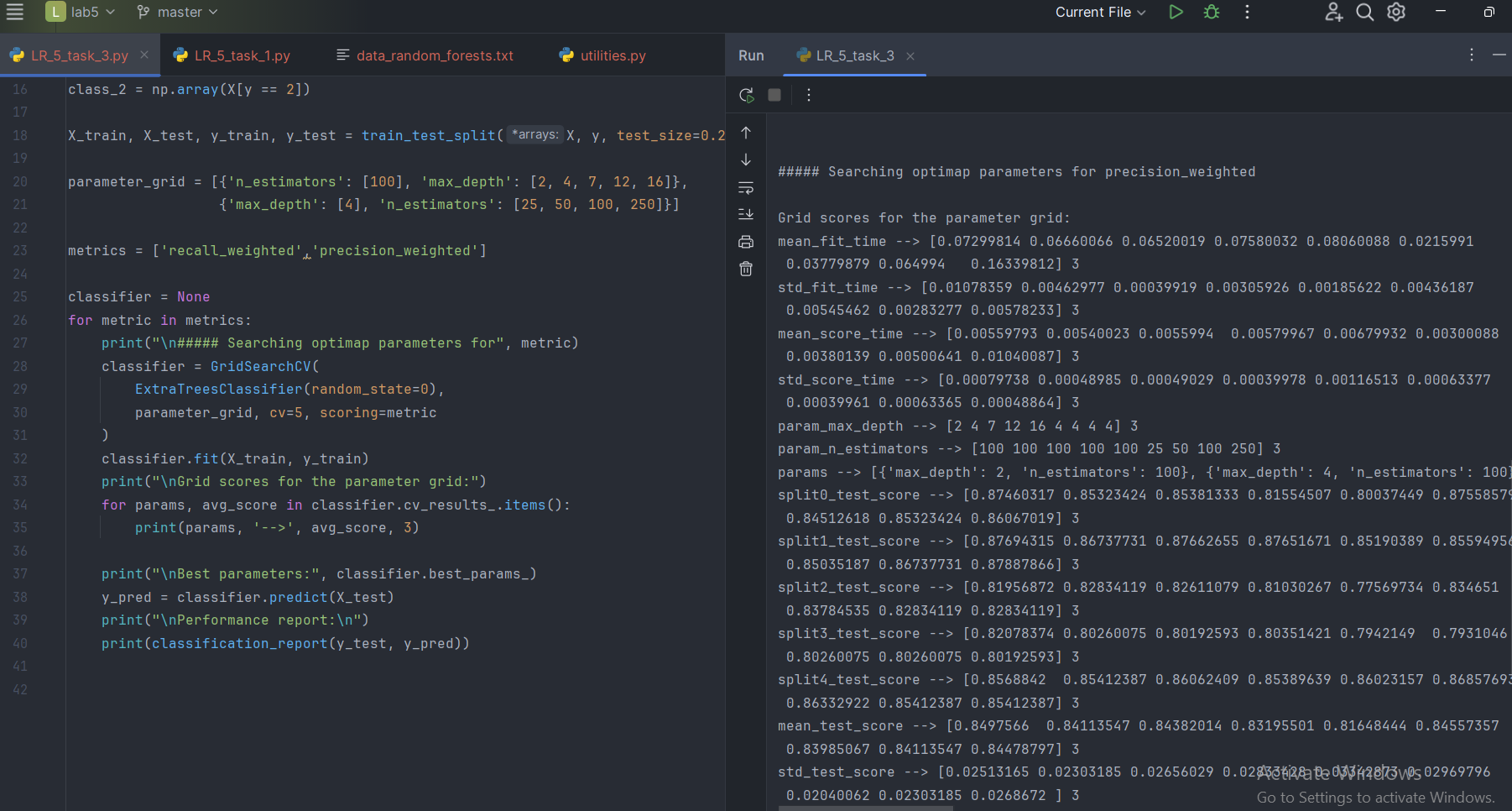


Рис 15. Результати пошуку оптимальних параметрів для precision

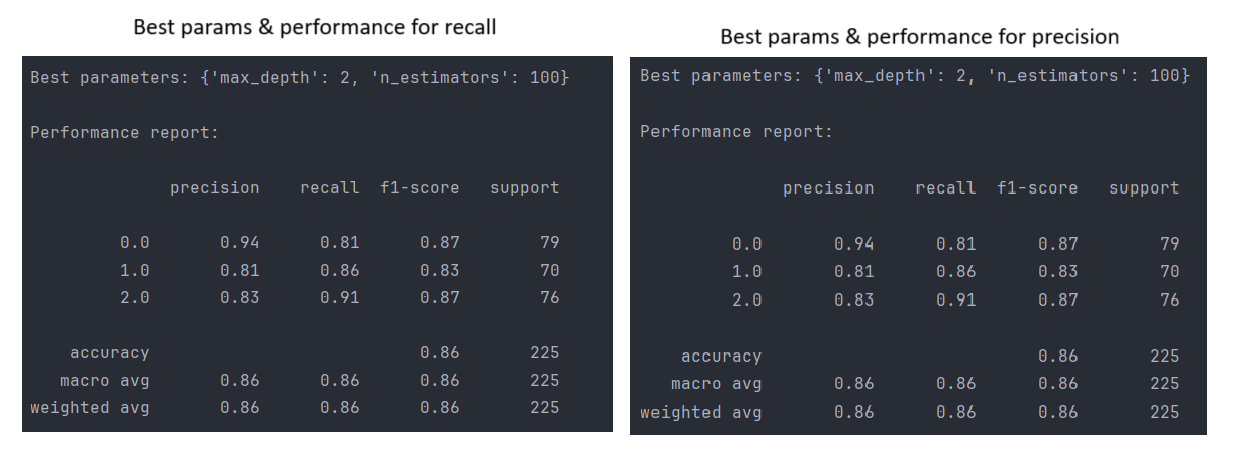


Рис 16. Найкращі параметри та оцінка продуктивності

**Висновок:** на даному етапі вдалося виявити, що найкращими оптимізуючими параметрами для precision & recall є max\_depth = 2 & n\_estimatirs = 100. Відповідно до оцінок продуктивності: найвищі показники має клас 0, тобто класифікатор найкраще його розпізнає.

**Завдання 4.** Обчислення відносної важливості ознак

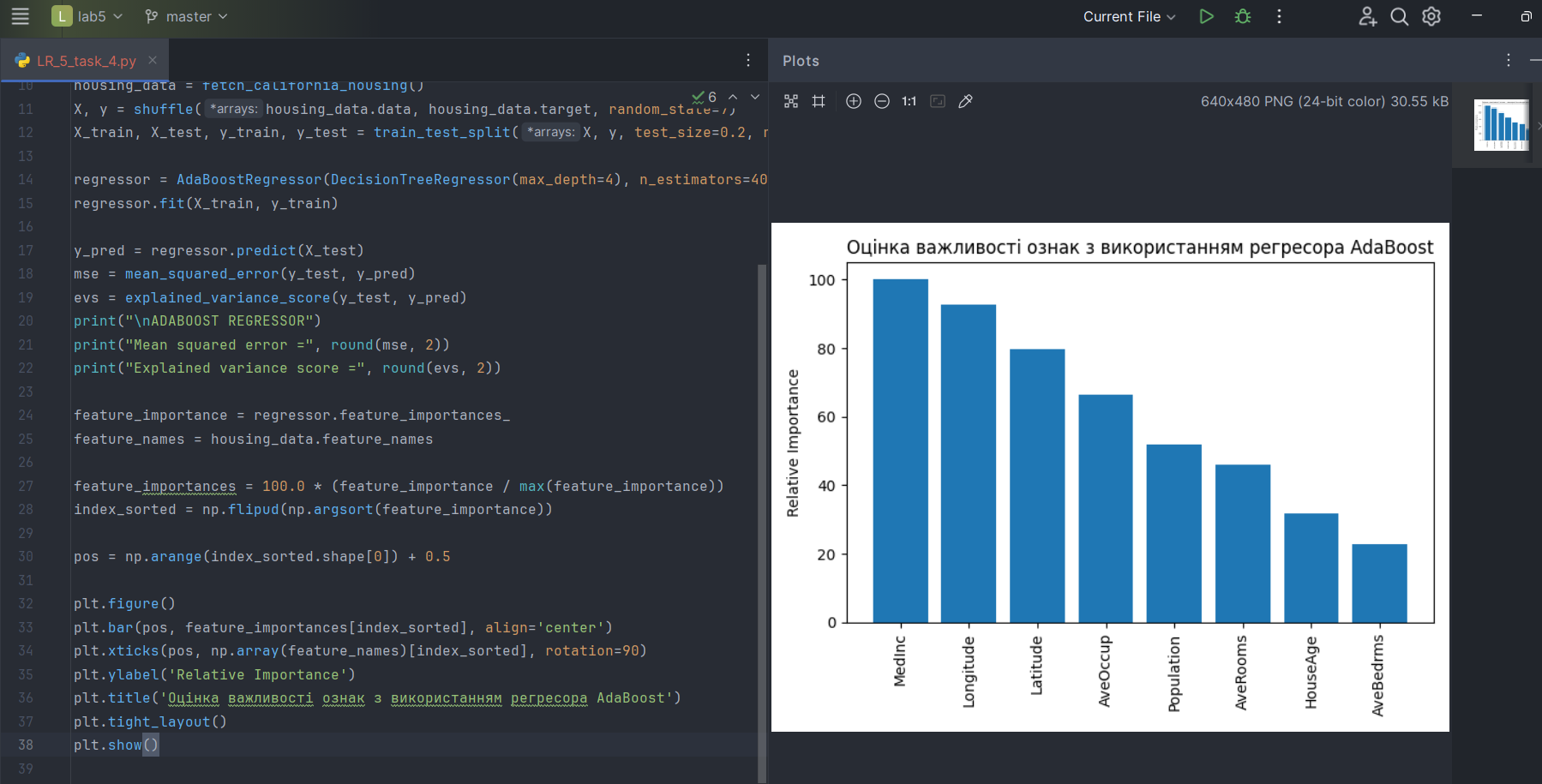
****

Рис 17. Графік

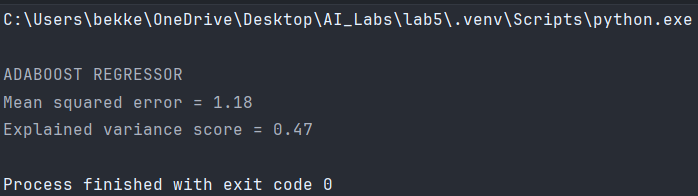


Рис 18. Метрики

**Висновок:** в результаті аналізу було виявлено, що найбільшу роль мають ознаки: medinc, longitude, latitude, тоді як останніми двома-трьома можна знехтувати. Відповідно до метрик: середньоквадратична помилка 1.18, якщо дані це просто цифри, є низькою, тобто відхилення не дуже велике, що є гарними результами. Дисперсія = 0.47 є середнім результом, що свідчить про те, що модель здатна пояснювати майже половину варіацій у даних.

**Завдання 5.** Прогнозування інтенсивності дорожнього руху за допомогою класифікатора на основі гранично випадкових лісів

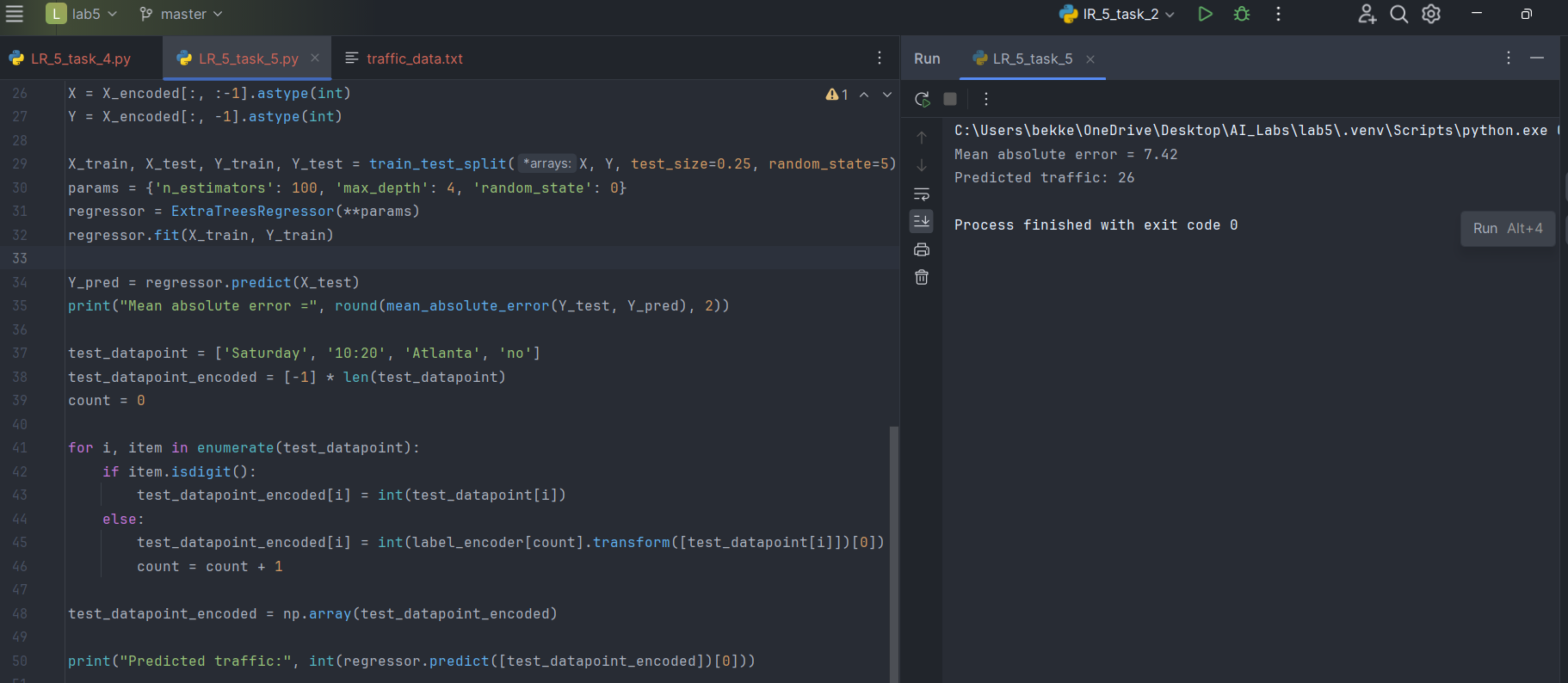
****

Рис 19. Метрики і прогноз

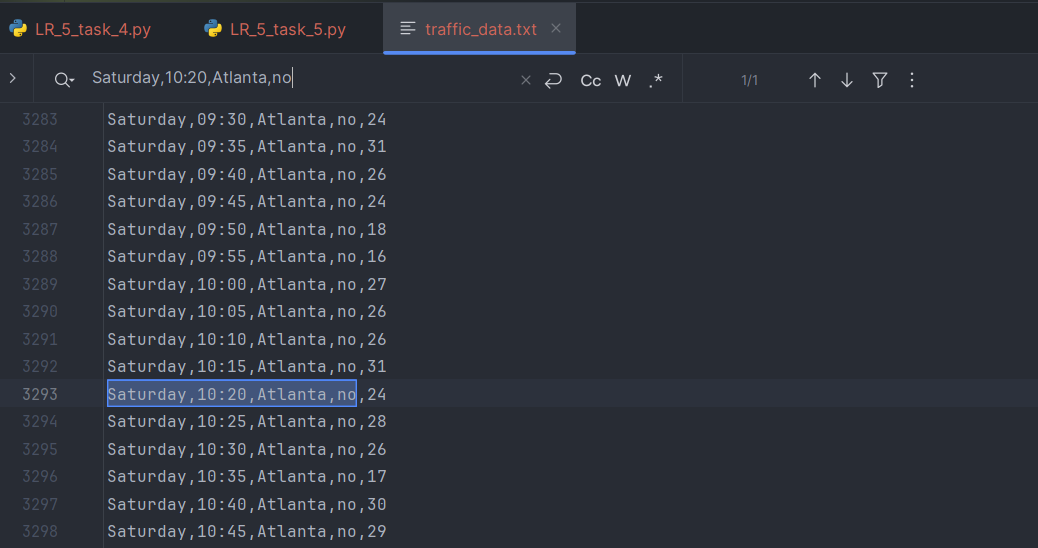


Рис 20. Актуальні дані

Оскільки прогноз дійсно близький до фактичного значення, можна зробити висновок, що ансамблеве навчання дійсно має гарні показники на таких наборах даних.

**Висновок:** на даній лабораторній роботі, я, за допомогою спеціалізованих бібліотек там мови програмування, навчився створювати та аналізувати класифікатори на основі випадкових та гранично випадкових лісів. Також отримав практичні навички з обробки дисбалансу класів, навчився знаходити оптимальні параметри за допомогою сіткового пошуку та обчислювати відносну важливість ознак.

Github: [link](https://github.com/Ajoke32/AILabs)