**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**ПОЛТАВСЬКА ПОЛІТЕХНІКА ІМЕНІ ЮРІЯ КОНДРАТЮКА**

**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА РОБОТОТЕХНІКИ**

Кафедра комп’ютерних та інформаційних технологій і систем

**Звіт з лабораторної роботи № 15**

з навчальної дисципліни

“Сучасні мови програмування”

**Виконав:**

*студент групи 401-ТН*

*Забільський Владислав Вадимович*

**Перевірив:**

*Ляхов Олександр Логвинович*

**Полтава**

**2024**

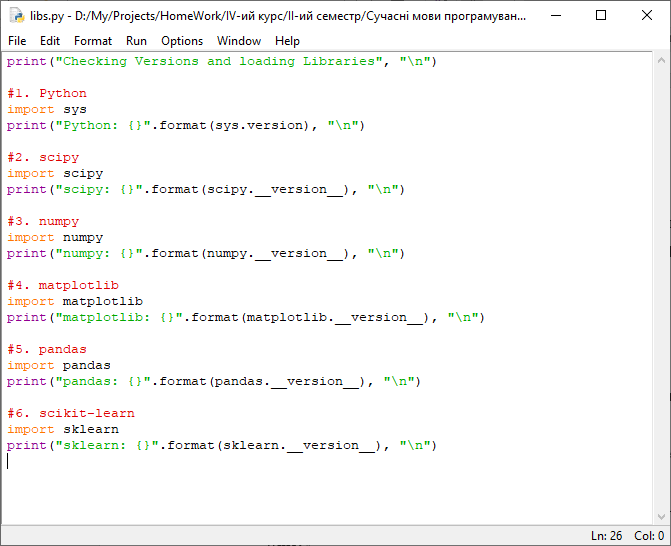
**Мета:** отримати навички виконання проекту вирішення задачі класифікації на даних, наданих замовником.

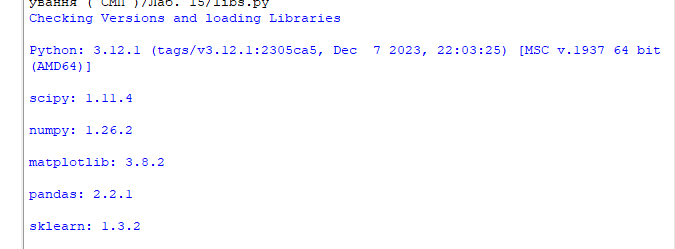
**Завдання:**

1. Поновити наведений код і отримати аналогічні результати.
2. Розв’язати задачу класифікації за даними про їх ознаки. Зробити для wine!

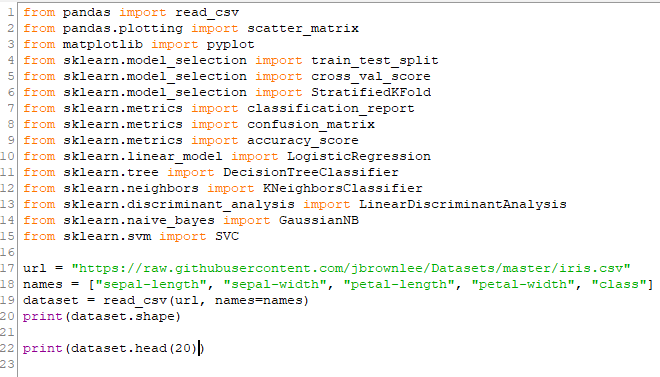
**Виконання завдання**

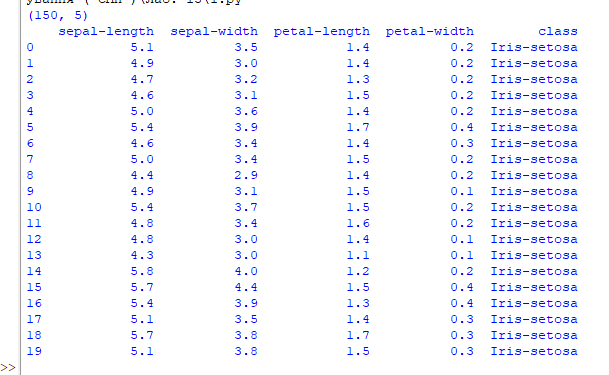
1. Поновити наведений код і отримати аналогічні результати.
2. Перевірка наявності та встановлення за необхідності ПЗ, а саме, платформі Python та бібліотек з набору SciPy.





1. Завантаження датасету.

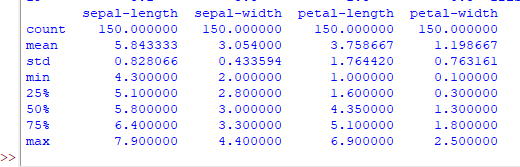


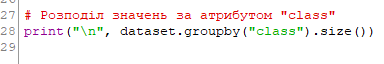


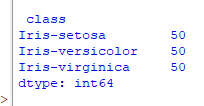
1. Аналіз датасету.

Отримуємо більш докладну інформацію про дані:



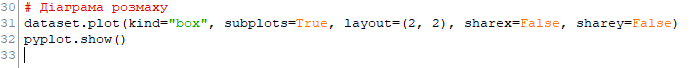


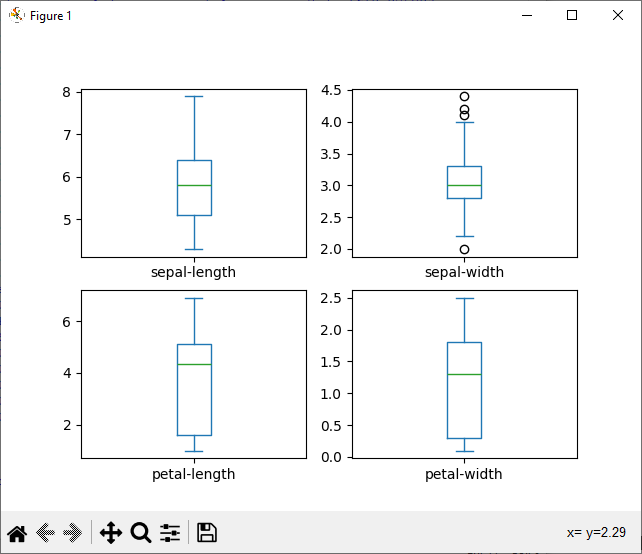


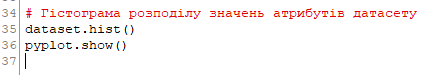


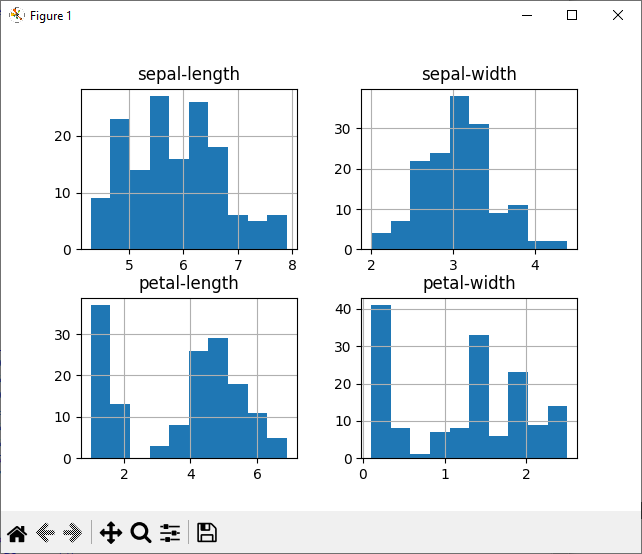
1. Візуалізація даних.

Одновимірні графіки:

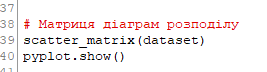


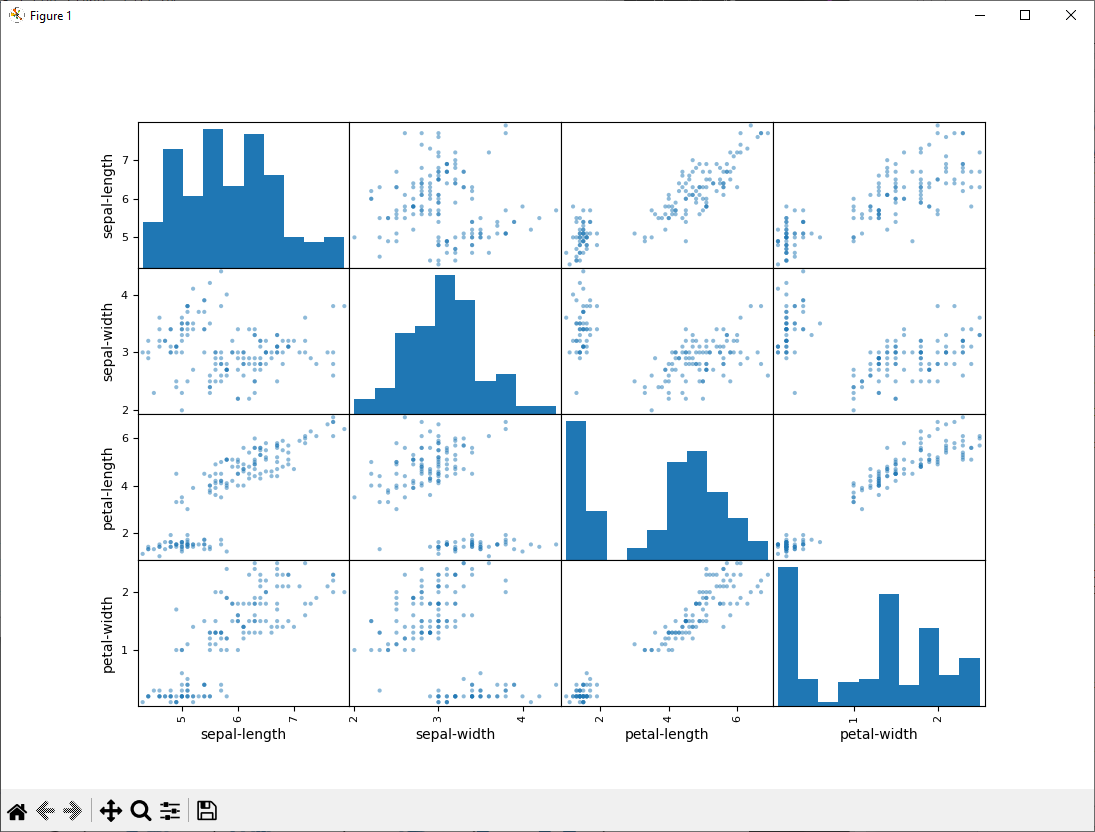






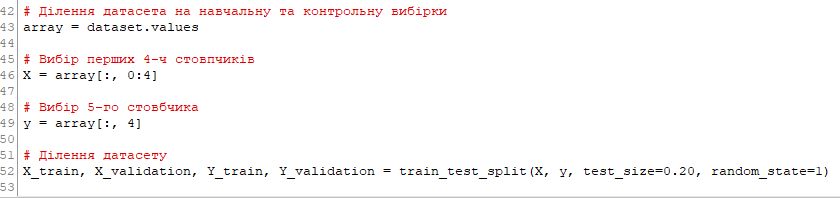
Багатовимірні графіки:



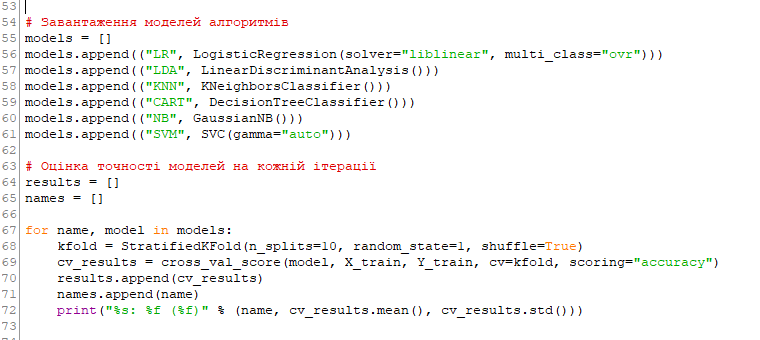


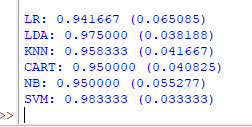
1. Вибір моделей алгоритмів для вирішення поставленої задачі.

Створення навчальної та тестової вибірки:

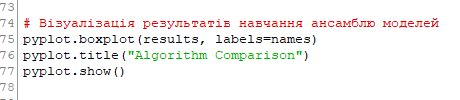


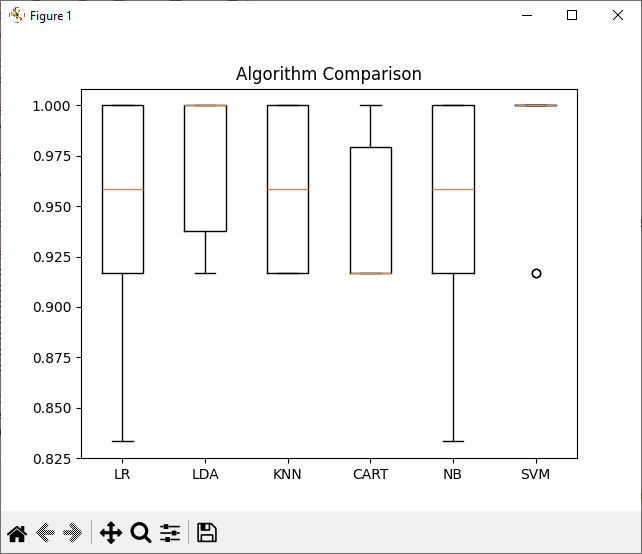
Ансамбль моделей:





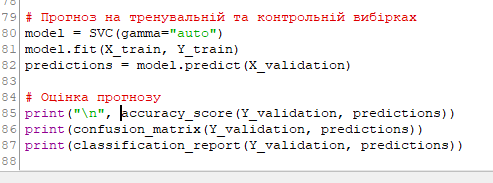
1. Навчання та вибір оптимального алгоритму

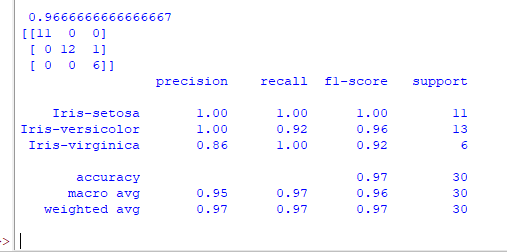




Червоний відрізок – середнє арифметичне значень точності всіх прогонів (10) моделі. Ящик та вуса – процентілі, які показують «розмах» значень точності на кожному прогоні.

1. Застосування навченого алгоритму для вирішення поставленої задачі.





1. Розв’язати задачу класифікації за даними про їх ознаки. Зробити для wine!
2. Перевірка наявності та встановлення за необхідності ПЗ, а саме, платформі Python та бібліотек з набору SciPy.

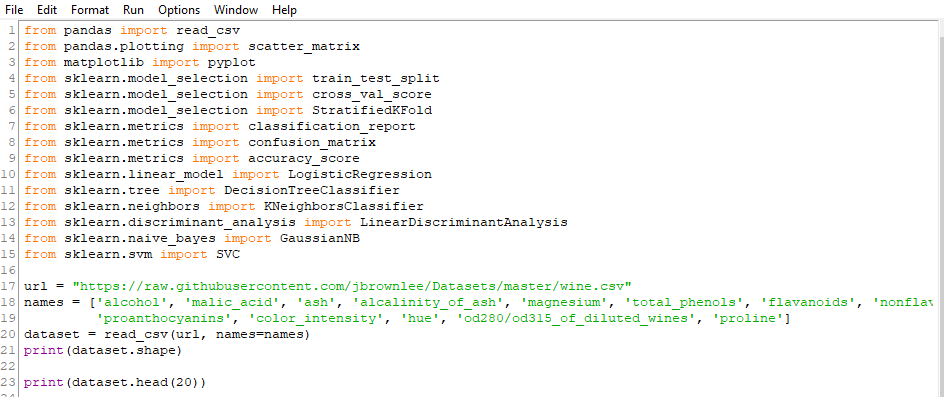
Перевірено у попередньому завданні.

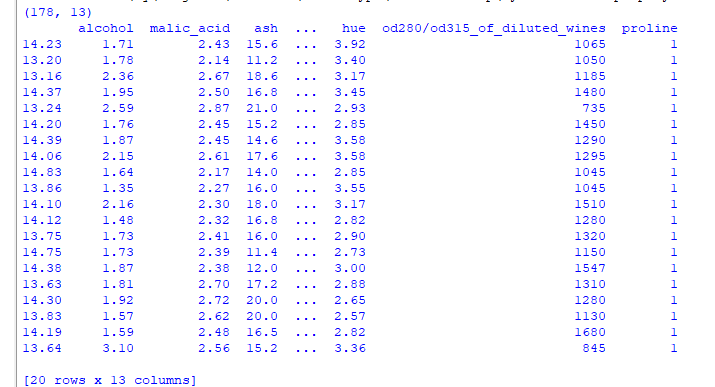
1. Завантаження датасету.

Необхідний датасет містить 178 спостережень.

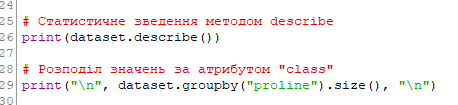
Використаємо це посилання для отримання даних про wine:

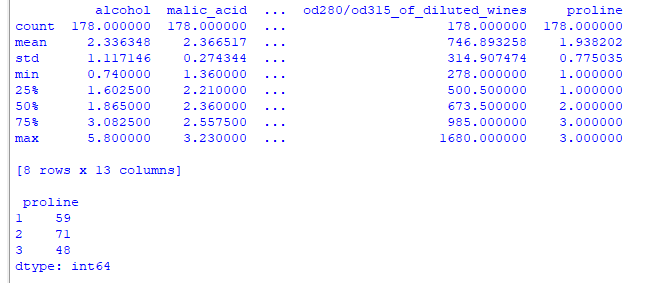
https://raw.githubusercontent.com/jbrownlee/Datasets/master/wine.csv





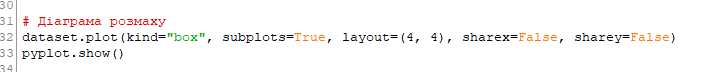
1. Аналіз датасету.

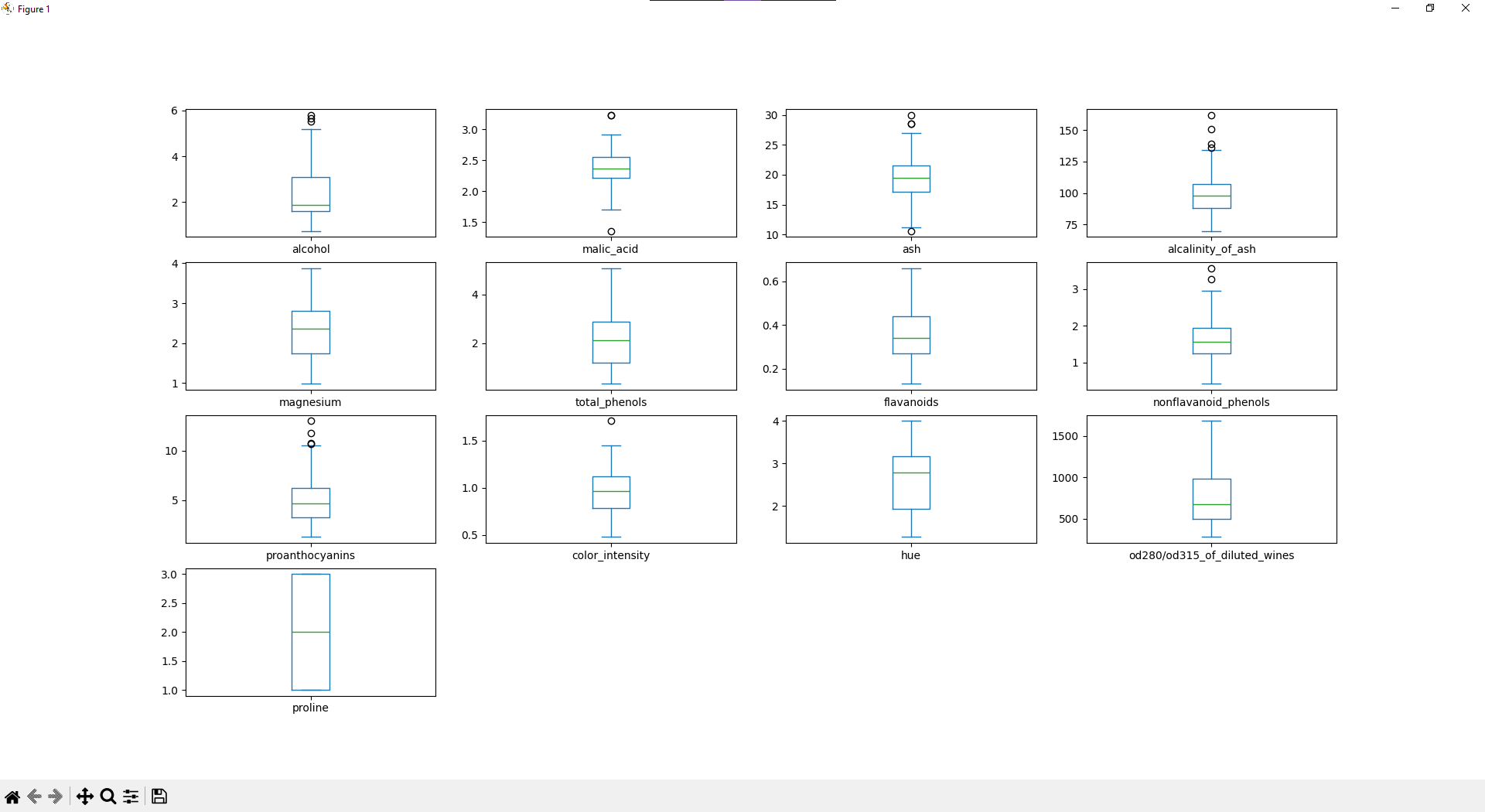


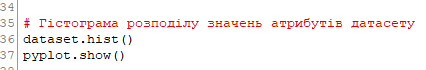


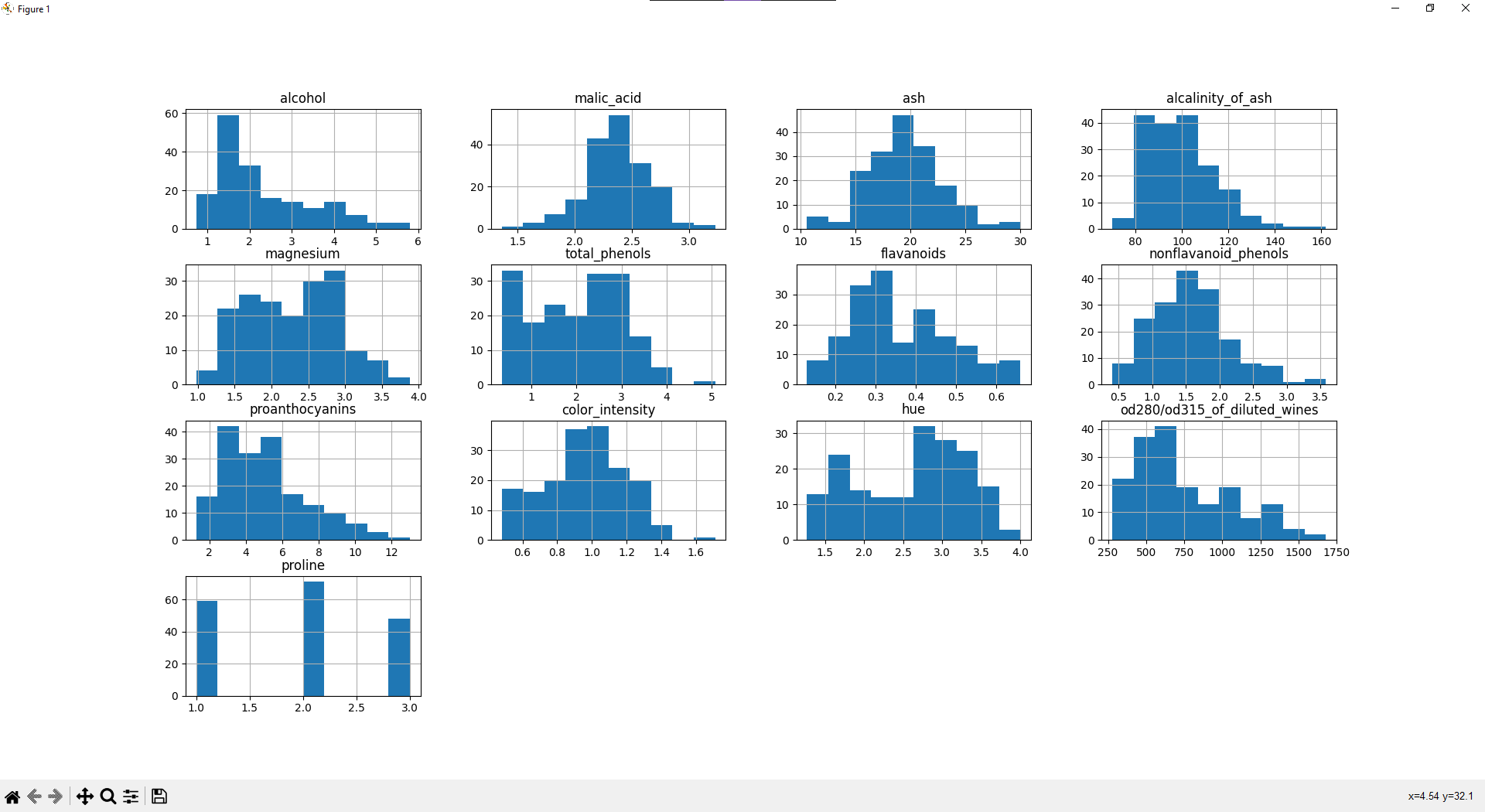
1. Візуалізація даних.

Одновимірні графіки:

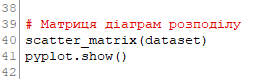


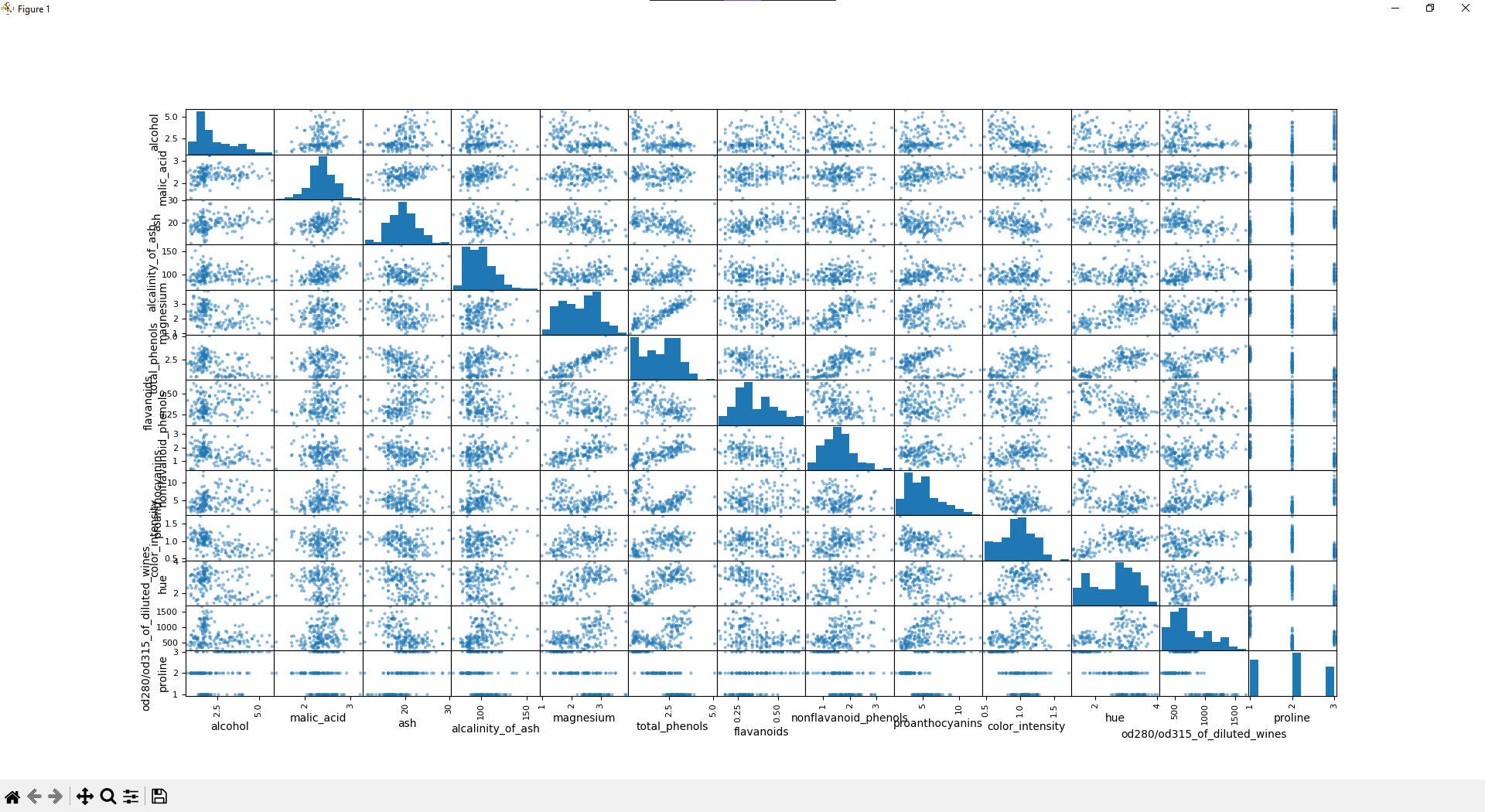






Багатовимірні графіки:



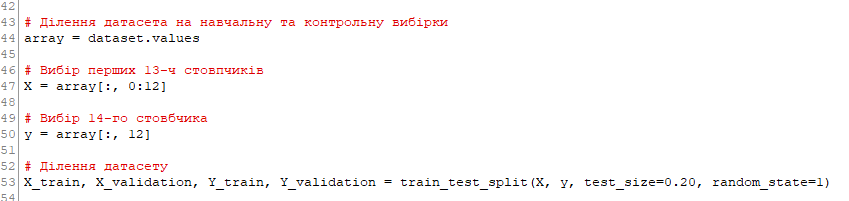


1. Вибір моделей алгоритмів для вирішення поставленої задачі.

Створення навчальної та тестової вибірки:

Поділимо завантажений датасет на два:

* 80% даних використаємо для навчання, оцінки і вибору найкращої серед відібраних моделей алгоритму розпізнавання (класифікації).
* 20% даних, які не будемо використовувати для навчання, а використаємо у якості тестового набору для оцінки якості кожної з моделей.

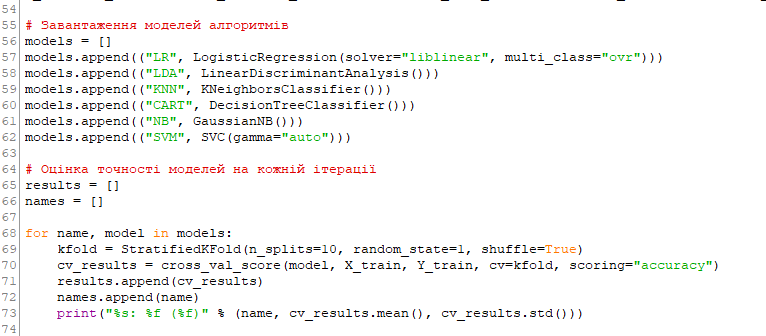


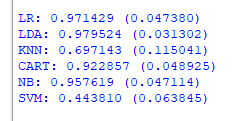
Так як не відомо, які моделі алгоритмів найкращі для наданого набору даних. Утворимо набір (ансамбль) алгоритмів і виберемо найкращий:

* Логістична регресія (LR)
* Лінійний дискримінантний аналіз (LDA)
* Метод k-найближчих сусідів (KNN)
* Класифікація й регресія за допомогою дерев (CART)
* Наївний байєсівський класифікатор (NB)
* Метод опорних векторів (SVM)

Залежність між ознаками може бути як лінійною, так і не лінійною. Отже ансамбль включає лінійні (LR и LDA) і нелінійні (KNN, CART, NB и SVM) моделі алгоритмів.

Ансамбль моделей:

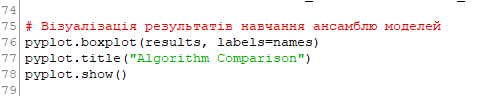


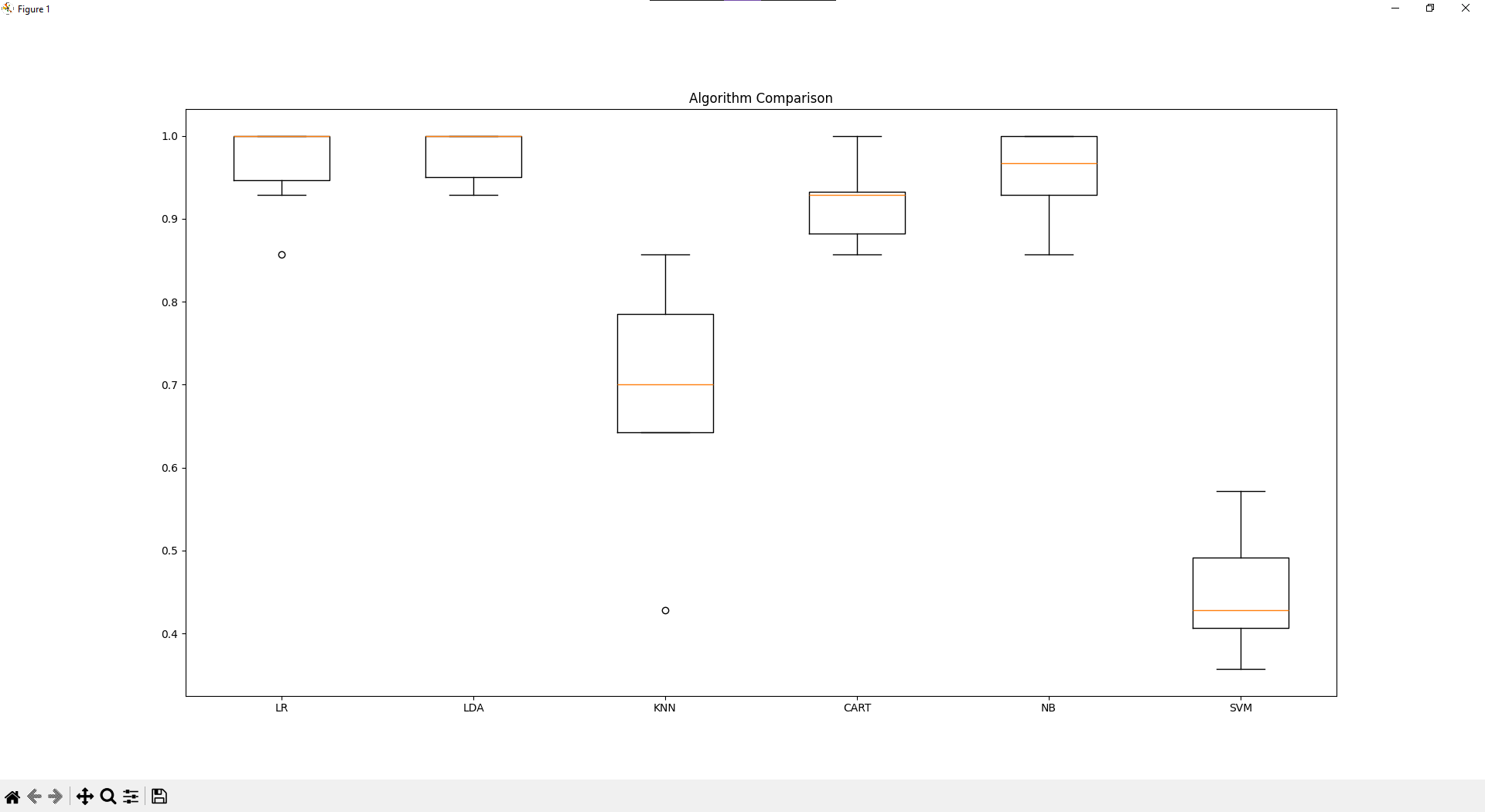


1. Навчання та вибір оптимального алгоритму

Висновок з наведених результатів, що найкращім є метод лінійного дискримінантного аналізу (LDA), оскільки його точність найбільша, біля 98% правильно класифікованих об’єктів з тренувального набору. Непоганою альтернативою може бути LR метод, але середнє квадратичне відхилення більше.

Візуалізуємо результати навчання ансамблю:

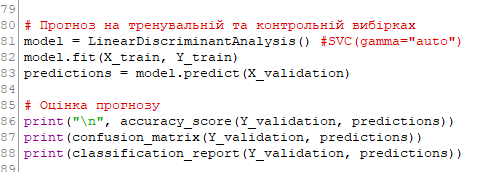


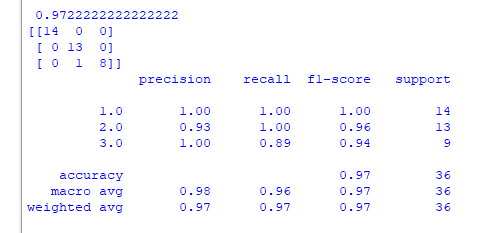


Червоний відрізок – середнє арифметичне значень точності всіх прогонів (10) моделі. Ящик та вуса – процентілі, які показують «розмах» значень точності на кожному прогоні.

1. Застосування навченого алгоритму для вирішення поставленої задачі.

Найкращий результат з ансамблю моделей має алгоритм LDA. У процесі його навчання не використовувалася контрольна (тестова) вибірка - 20% даних з датасету. Їх ми й використаємо для валідації алгоритму.





Дані для тестування містять розміри 36-ти об’єктів (20%), що не використовувалися при навчанні. Точність прогнозу на цих даних 97%. Це відповідає точності на тренувальному наборі 98%. Матриця кількості помилок дає уявлення про одну помилку при розпізнаванні належності до класу № 3.