ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ №2

за курсом "Python для Data Scientist"

студента/студентки групи ПА-

HarryJamesPotter/HermioneJeanGranger

кафедра комп’ютерних технологій, ДНУ

2024/2025

Тема: «Побудова матриці кореляції та виявлення факторних ознак»

Постановка задачі:

Ознаки, які мають найбільший вплив на класифікацію в наборі даних, зазвичай називаються **важливими ознаками** (important features) або **факторами, що визначають класифікацію**. Вони можуть бути виявлені за допомогою різних методів, таких як:

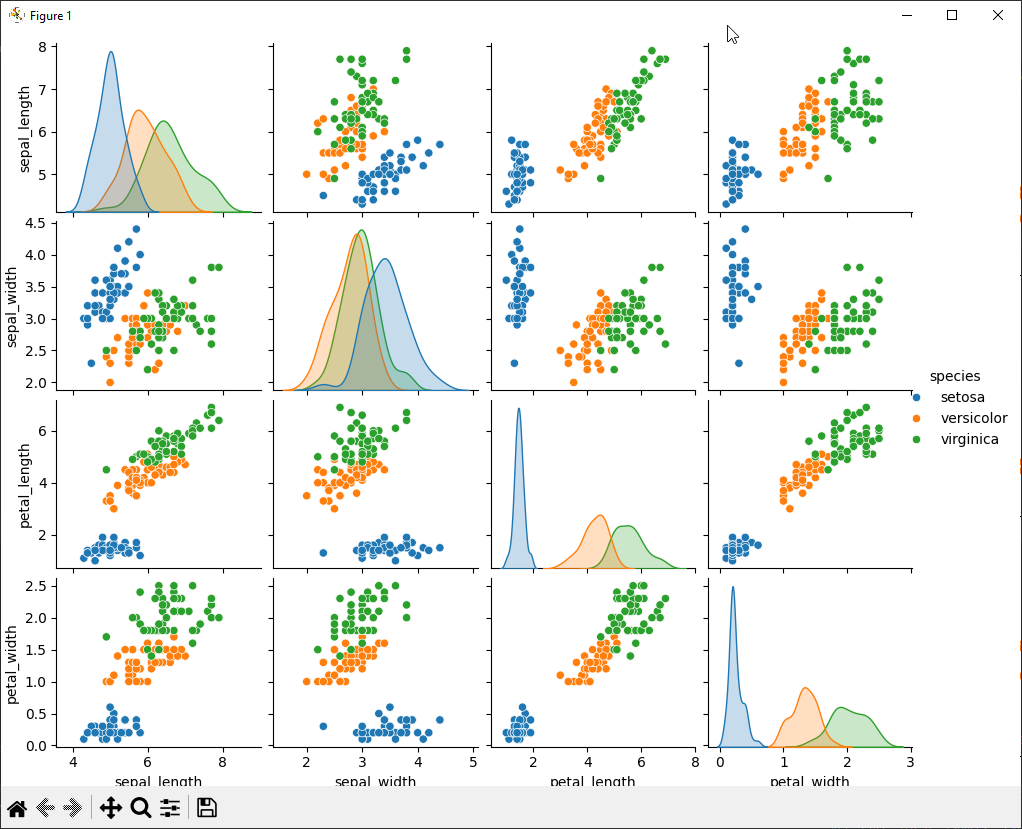
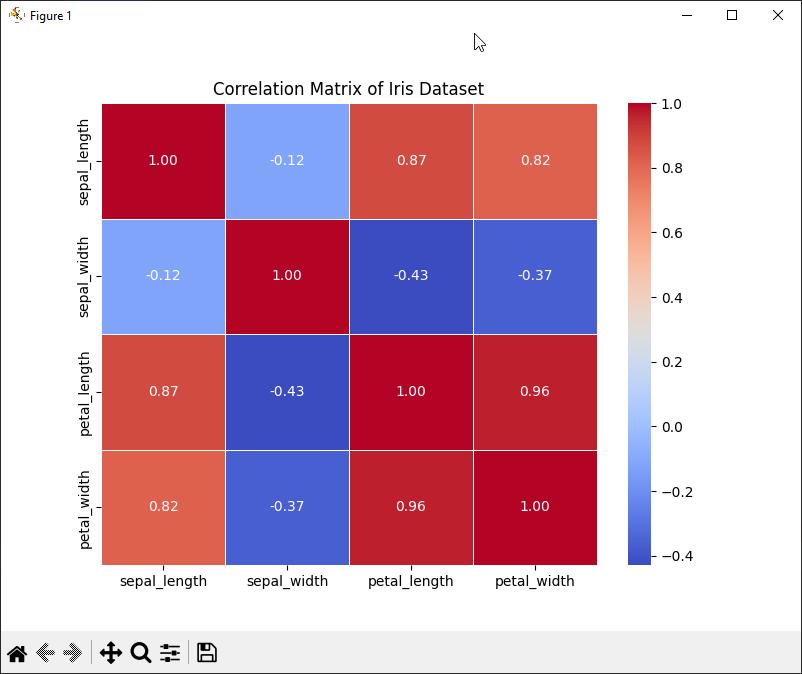
1. **Аналіз кореляції** – визначення, як сильно кожна ознака пов'язана з класом або іншими ознаками.
2. **Методи відбору ознак** (Feature Selection) – алгоритми, що оцінюють важливість кожної ознаки для покращення точності моделі.
3. **Методи регресії** – виявлення коефіцієнтів або ваг для кожної ознаки, що показують її значущість.
4. **Деревоподібні методи** (наприклад, Random Forest або Gradient Boosting) – автоматично визначають важливість ознак шляхом аналізу дерева рішень.
5. **Інші**

Обрати **ОДИН**, будь-який метод.

Проаналізувати **ОДИН** з наборів даних:

* Іриси Фішера: <https://www.kaggle.com/datasets/uciml/iris>
* Пінгвіни Пальмера: <https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data>
* Набір даних полювання на гриби: <https://www.kaggle.com/datasets/uciml/mushroom-classification>
* <https://www.kaggle.com/datasets/erdemtaha/cancer-data/data>
* <https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset>
* <https://www.kaggle.com/datasets/reihanenamdari/breast-cancer>
* Чи інший який ви знаєте ☺

Приклад (Іриси Фішера)

Бібліотеки

* <https://seaborn.pydata.org/>
* Pandas
* Matplotlib

Посилання:

* <https://www.kaggle.com/>
* <https://khashtamov.com/ru/pandas-introduction/>
* <https://habr.com/ru/companies/otus/articles/741064/>
* <https://www.datacamp.com/blog/what-is-data-science-the-definitive-guide>