Зображення, що містить текст

Автоматично згенерований опис

Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут імені Ігоря Сікорського”

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

**Лабораторна робота №7**

Аналіз даних з використанням мови Python

**Тема:** Кластеризація та регресія в scikit-learn

**Варіант:** 1

Виконав Перевірив:

студент групи ІП-11: Тимофєєва Ю. С

Панченко С. В.

Київ 2023

Зміст

[1 Мета лабораторної роботи 6](#__RefHeading___Toc161_232900929)

[2 Завдання 7](#__RefHeading___Toc163_232900929)

[3 Виконання 8](#__RefHeading___Toc165_232900929)

[3.1 Перетворення даних 8](#__RefHeading___Toc167_232900929)

[3.2 Навчання та тестування моделі 10](#__RefHeading___Toc169_232900929)

[3.3 Зменшення вимірів (Dimensionality reduction) 11](#__RefHeading___Toc171_232900929)

[4 Висновок 13](#__RefHeading___Toc173_232900929)

# Мета лабораторної роботи

Ознайомитись з побудовою моделей для вирішення задач регресії та кластеризації в scikit-learn, визначити основні оцінки цих моделей.

# Завдання

Написати програму,яка здійснює попередню обробку даних, навчає та тестує(при потребі) модель, що виконує завдання відповідно до варіанту, оцінити модель за допомогою відповідних метрик.

Оформити звіт. Звіт повинен містити:

—титульний лист;

—код програми;

—результати виконання коду.

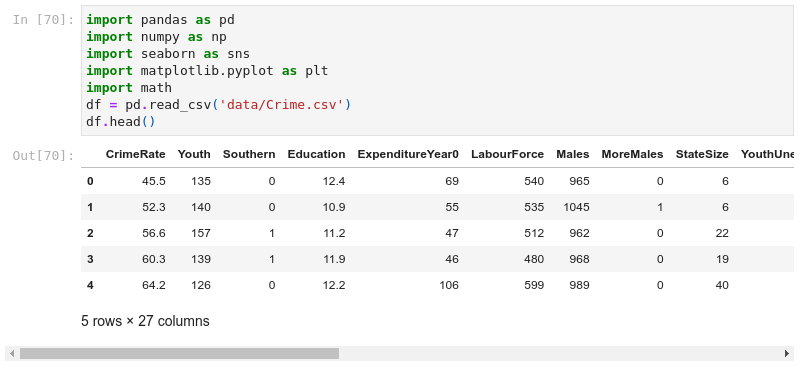
Продемонструвати роботу програми та відповісти на питання стосовно теоретичних відомостей та роботи програми.

Варіант 1: lab2/Crime.csv. Використати будь-яку комбінацію незалежних ознак (не менше двох), щоб спрогнозувати рівень злочинності в даний час.

# Виконання

## Перетворення даних

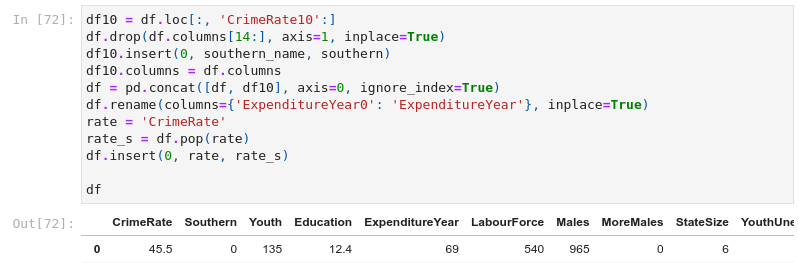
Для початку імпортуємо модулі pandas, numpy, seaborn, matplotlib. Завантажимо датафрейм.

  
  
Рисунок 3.1 - Завантаження датафрейму

Переставимо колонку у "Southern" на перше місце, оскільки розташування штату не змінювався.

  
  
Рисунок 3.2 - Переставлення колонки

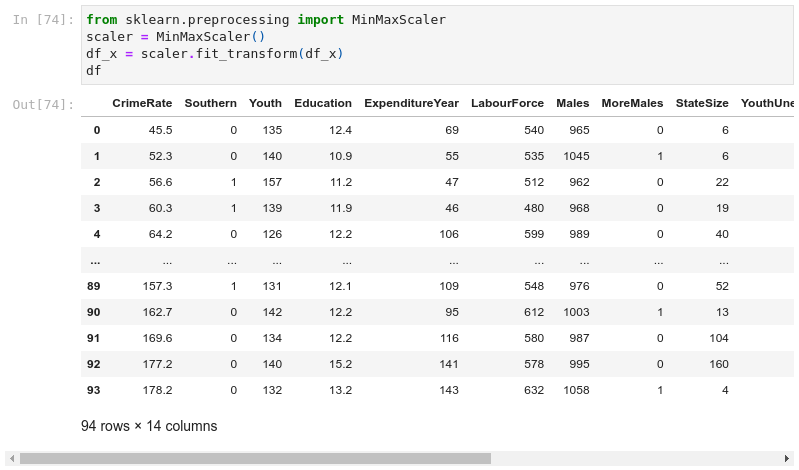
Переставимо показники через 10 років у звичайні колонки, розділивши датафрейм надвоє та з'єднавши частини вертикально. Таким чином ми збільшуємо розмір вибірки у два рази, тому модель матиме більше даних, аніж у тому випадку якби ми знаходили, наприклад, середнє арифметичне між колонками за 10 років.

  
  
Рисунок 3.3 - Перетворення даних

Розділимо дані на аргументи та значення.

  
  
Рисунок 3.4 - Розділення даних на аргументи та значення

Зробимо масштабування даних за допомогою MinMaxScaler.

  
  
Рисунок 3.5 - Масштабування даних

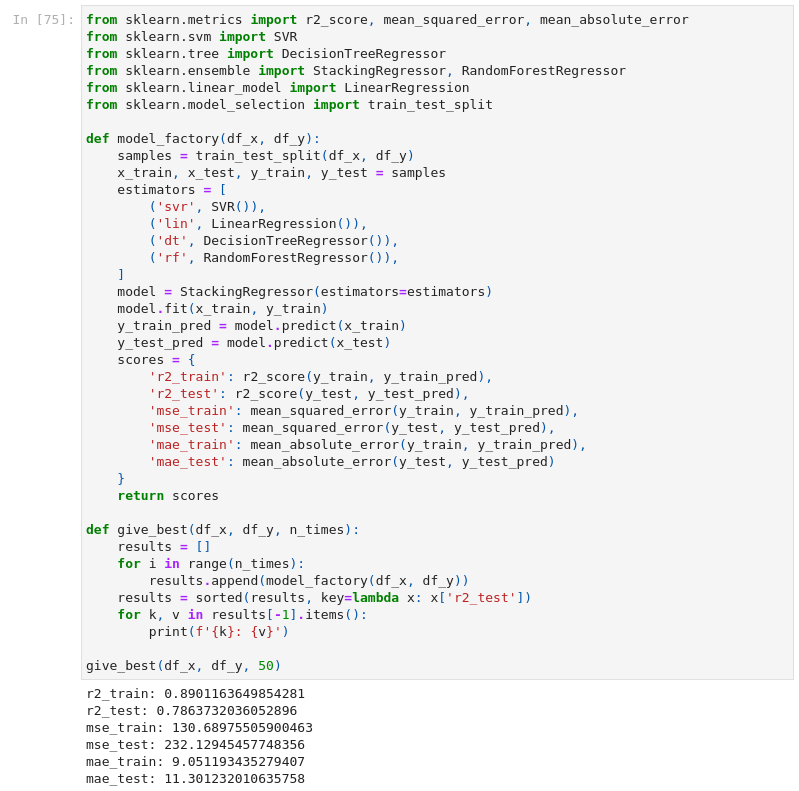
## Навчання та тестування моделі

Зараз побудуємо модель, де будемо використовувати одночасно декілька алгоритмів для того, щоб мати кращий результат. До прикладу, оберемо SVR, LinearRegression, DecisionTree та RandomForest. Для цього імпортуємо з пакету sklearn.ensemble StackingClassifier та RandomForestRegressor, з sklearn.svr - SVR, з sklearn.linear\_model - LinearRegression, з sklearn.tree - DecisionTreeRegressor.

Напишемо функцію model\_factory та проведемо тренування n-кількість разів, зберігатимемо найкращу з моделей. Також будемо всередині використовувати функцію train\_test\_split, яка ділить датасет на 75% тренувальних та 25% тестових даних за замовчуванням.

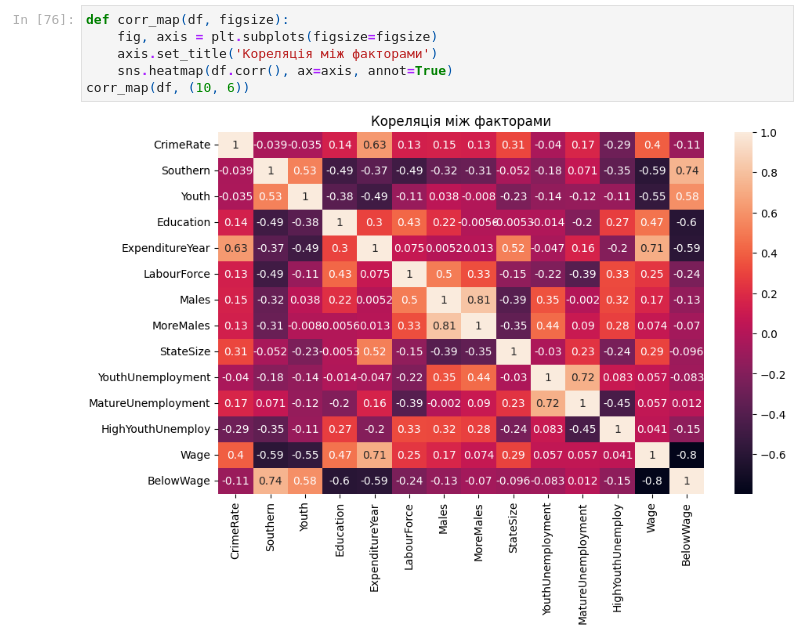
Оцінювання буде проводитися параметр R2, або коефіцієнт детермінації, який кількісно визначає частку дисперсії залежної змінної, яку можна передбачити на основі незалежних змінних. Також використаємо середню квадратичну похибку та середню абсолютну.

Проведемо певну кількість тренувань моделі та оберемо найкращу.

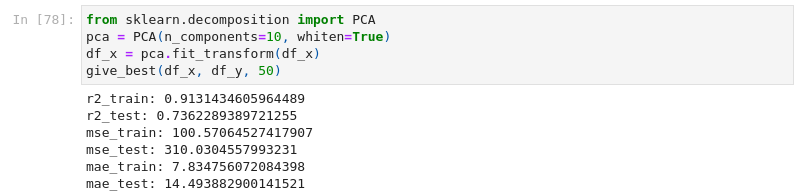
  
  
Рисунок 3.6 - Тренування моделей

## Зменшення вимірів (Dimensionality reduction)

Однак дана модель має один мінус. У ній забагато даних обробляється. І, можливо, результати можна покращити, зробивши dimesionality reduction. Для початку побудуємо матрицю кореляцій.

  
  
Рисунок 3.7 - Матриця кореляцій

Як бачимо, матриця має доволі багато взаємозалежних величин, тому зменшимо їхню кількість. Для цього використаємо PCA - Principal Component Analysis, або метод головних компонент. В Імпортуємо відповідний пакет PCA з sklearn.decomposition.

  
  
Рисунок 3.8 - Зменшення залежних величин

Як бачимо, точність для тестових даних погіршилася, а тому сенсу зменшувати далі нема, оскільки величини сильно взаємозалежні.

# Висновок

Під час виконання даної лабораторної роботи я ознайомився з побудовою моделей для вирішення задач регресії та кластеризації в scikit-learn, визначив основні оцінки цих моделей.

По-перше, перетворив дані: перемістив інформацію зі стовпчиків, яка була зафіксована через 10 років, у звичайні стовпчики, таким чином зменшивши їхню кількість.

По-друге, зробив масштабування аргументів за допомогою MinMaxScaler.

По-третє, спроектував багатошарову модель з використанням таких алгоритмів як: SVR, DecisionTreeRegressor, RandomForestRegressor, LinearRegression.

По-четверте, розділив дані на навчальні та тестові у відношенні до основного датасету як 75% і 25% відповідно.

По-п’яте, для оцінки якості моделі використав параметр R2, середню квадратичну та абсолютну похибки.

По-шосте, побудував матрицю кореляцій та спробував зменшити кількість стовпчиків за допомогою PCA. Однак результати виявилися гіршими ніж з використанням усіх даних.