# Практична робота № 14

## Тема: Регресія в Python

***Мета:*** набути практичні навички роботи в середовищі розробки Python та проведенні найпростішого аналізу даних.

## Література

*Документація по бібліотеці Seaborn -* [*https://seaborn.pydata.org*](https://seaborn.pydata.org/) *seaborn*.*pairplot()- https://seaborn.pydata.org/generated/seaborn.pairplot.html seaborn.boxplot() - https://seaborn.pydata.org/generated/seaborn.boxplot.html Statsmodels* [*https://www.statsmodels.org/stable/index.html*](https://www.statsmodels.org/stable/index.html)

# Зміст роботи

За допомогою машинного навчання можна прогнозувати подальші події шляхом аналізу попереднього досвіду . Наприклад, скласти прогноз погоди на завтра, або вгадати курс акцій на біржі, або діагностувати хворобу пацієнта, ґрунтуючись на його попередньої історії хвороби.

Машинне навчання

Регресія

Класифікація

Неконтрольоване навчання

Контрольоване навчання

*класи алгоритмів*

Класифікація може визначити категорію вхідних даних або наявність, або відсутність якоїсь їх особливості. Наприклад, намагатися розпізнати написану цифру або визначити, чи міститься на зображенні кіт.

Регресія ж обчислює певне число або вектор - наприклад, завтрашню температуру або ціну на акції Google.

Лінійна регресія (Linear regression) - модель залежності змінної x від однієї або декількох інших змінних (факторів, регресорів, незалежних змінних) з лінійною функцією залежності. Лінійна регресія відноситься до задачі визначення «лінії максимальної відповідності умовам» через набір точок даних і стала простим попередником нелінійних методів, які використовують для навчання нейронних мереж.

***Завдання 1.*** Дослідити залежність продажів від витрат на рекламу на телебаченні, радіо та в газеті. Вхідні дані взяти з файлу Advertising.txt.

***Завдання 1.1***. Завантажити набір даних.Роздрукувати перших 5 рядків отриманого датафрейму

# *Завдання 1.2.* Побудувати парні графіки

Для того, що б наочно побачити можливу статистичну залежність в даних необхідно побудувати парні графіки. Зробити це зручно за допомогою бібліотеки *seaborn*в якій є метод *pairplot* який будує попарні залежності ознак з датасету.

Ознайомитися з параметрами методу:

sns.pairplot(df, palette='dict' , x\_vars=('TV', 'Radio', 'Newspaper','Sales '), y\_vars=('TV', 'Radio', 'Newspaper','Sales'), kind='scatter', diag\_kind

='auto')

Варіанти результату:



На діагоналі представлено розподіл відповідної ознаки, наприклад, скільки всього було Sales (продаж).

# Завдання 2. Побудувати діаграму розсіювання (TV i sales, radio i sales, newspaper i sales) для візуалізації зв’язку.

# Приклад коду:

fig, axs = plt.subplots(1, 3, sharey=True)

df.plot(kind='scatter', x='TV', y='Sales', ax=axs[0], figsize=(16, 8)) df.plot(kind='scatter', x='Radio', y='Sales', color='red', ax=axs[1]) df.plot(kind='scatter', x='Newspaper', y='Sales', color='green', ax=axs[2])

Результат:



## Питання, що виникають в процесі аналізу і потребують відповідей:

* *Чи існує взаємозв’язок між рекламою в газетах, на радіо і ТV та продажами? Наскільки сильні зв’язки?*
* *Яка реклама сприяє продажам?*
* *Який вплив має кожен тип реклами на продажі?*
* *Враховуючи витрати на рекламу на певному ринку, чи можна прогнозувати продаж?*

З графіків можна зробити кілька висновків за даними, щодо того, як впливає реклама в газетах, радіо і ТV на продажі. Видно, що найменший вплив на продажі від реклами в газетах, потім на радіо і нарешті найбільший від реклами на ТV.

# *Завдання 3.* Розрахувати коефіцієнт кореляції даних.

Для розрахунку кореляції використаємо метод **corr()** - за замовчуванням це кореляції Пірсона.

Результат:



Коефіцієнт кореляції між рекламою на ТV і продажами = 0,782224 (78 відсотків), далі йде радіо - 0,576223 (57%) ну і нарешті газети – 0,228299 (22,8%).

*Отже, розрахований коефіцієнт кореляції свідчить про наявність значного зв'язку між рекламою на ТV і продажами*

***Завдання 3.*** Провести додатковий аналіз:

Побудуйте лінію регресії для продажів в залежності від реклами на TV, Radio і Newspaper. за допомогою функції (sns.pairplot()), і 95% довірчий інтервал для цієї регресії: y ~ x

Результат:



Для зміни кольору лінії регресії використайте наступний параметр

plot\_kws={'line\_kws':{'color':'red'}}

***Завдання 4.*** Побудувати тривимірний графік. Точки різних категорій відобразити різними фігурами.