МІНІСТЕРСТВО ОСВІТИ І НАУКИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Кафедра систем штучного інтелекту



Лабораторна робота №6

З курсу “Візуалізація даних”

Виконав:  
студент групи КН-310

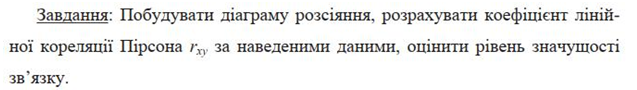
Бурак Марко

Викладач:

Виклюк Я.І.

Львів – 2020

**Мета:** оцінка лінійного кореляційного зв’язку





Пояснення: кожен студент повинен до кожної цифри у клітинках добавити останню цифру порядкового номера у журналі, але з певними корективами: 307 – (+1), 308 – (+2), 309 – (-1), 310 – (-2).

Мій порядковий номер в журналі – 4, остання цифра – 4, група – КН-310.

Код програми:

**Імпортую бібліотеки для роботи з датасетом**

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

from sklearn.linear\_model import LinearRegression

**Додаю до кожного значення з датасету 2, як сказано за умовою задачі**

correct\_number = 4 - 2

**Код програми**

data = pd.read\_csv("data.csv", delimiter=";")

data["X"] += correct\_number

data["Y"] += correct\_number

data.plot.scatter("X", "Y")

corr = np.corrcoef(data["X"], data["Y"])[0][1]

print("Coef Correlation:", round(corr, 5))

analys\_corr(corr)

m, b = np.polyfit(data["X"], data["Y"], 1)

plt.plot(data["X"], m\*data["X"] + b, color="green")

plt.show()

**У цій функції я формую датасет з даних data,**

**З полями X та Y**

**Ця частина коду, відповідає за аналіз кореляції**

def analys\_corr(corr):

if corr == 1:

print("A perfect positive relationship.")

elif 0.8 <= corr < 1:

print("A fairly strong positive relationship.")

elif 0.6 <= corr < 0.8:

print("A moderate positive relationship.")

elif 0 < corr < 0.6:

print("A weak positive relationship.")

elif corr == 0:

print("No relationship.")

elif -0.6 < corr < 0:

print("A weak negative relationship.")

elif -0.8 < corr <= -0.6:

print("A moderate negative relationship.")

elif -1 < corr <= -0.8:

print("A fairly strong negative relationship.")

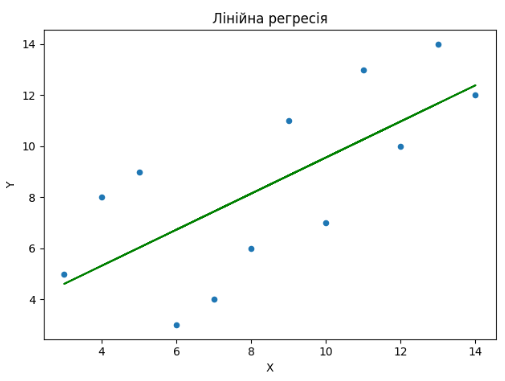
elif corr == -1:

print("A perfect negative relationship.")

else:

print("Cannot analys coef corr!")

**Після апробації програми отримаємо такий графік:**



На цьому графіку бачимо відношення значень X та Y.

Назва графіку – це “Лінійна регресія”

Також тут вказано, точкове представлення датасету, яке позначено синім кольором.

Зеленим ж кольором позначено лінійну регресію.

Значення з датасету, розміщені хаотично відносно прямої, це можна бачити на графіку, деякі значення знаходяться близько до прямої, а деякі навпаки далеко.

Після апробації програми також отримуємо вивід у консоль.



Перше значення відповідає за кореляцію.

А друге аналіз цієї кореляції.

Кореляції між двома змінними X та Y, набуває значень від −1 до +1 включно.

Коефіцієнт кореляції між двома змінними дорівнює коваріації двох змінних, або сумі добутків відхилень, поділеній на добуток їх стандартних відхилень.

Спочатку проаналізуємо всі можливі випадки.

Значення +1 означає, що залежність є лінійною і всі точки лежать на прямій лінійної регресії. Також відображає зростання Y при зростанні X.

Значення -1 означає, те ж що і +1, проте відображає Y при зростанні X.

Найгіршим варіантом, хаотичне розсташування точок, відбувається при значенні кореляції 0.

З вище згаданих фактів про кореляцію, можна зробити висновок, що якщо значення кореляції наближається до 0, тоді точки розсташовані подалі від лінійної регресії. З іншої ж сторони, при наближені до 1, значення приближаються до прямої регресії.

У нашому ж випадку, значення кореляції рівне 0.70629, це число відповідає за те, що точки не лежать на самій прямій регресії, проте, і не розсіюються хаотично. Це хороший показник, але далекий до ідеального.

**Висновок:** Я навчився будувати та візуалізувати діаграму розсіювання, також розрахував коефіцієнт Пірсона за навединими даними, також провів внвліз цієї кореляції, для того, щоб з’ясувати як розсташовані точки на площині відносно прямої регресії.