

**Звіт**  
про виконання лабораторної роботи № 1  
«Статистичний та кластерний аналіз»

Виконав:  
студент групи 126-32/1  
Остріков Д.

Перевірив:  
Рибак А. В.

# Лабораторна робота № 1

## «Статистичний та кластерний аналіз»

### Мета:

Виконати кластерний аналіз для встановлення зв'язку між об'єктами та провести статистичний аналіз для кожного сформованого кластеру

### Хід роботи:

1. Створив Virtual Environment для проєкту, інсталиював та імпортував потрібні бібліотеки (pandas, matplotlib, sklearn).

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs
```

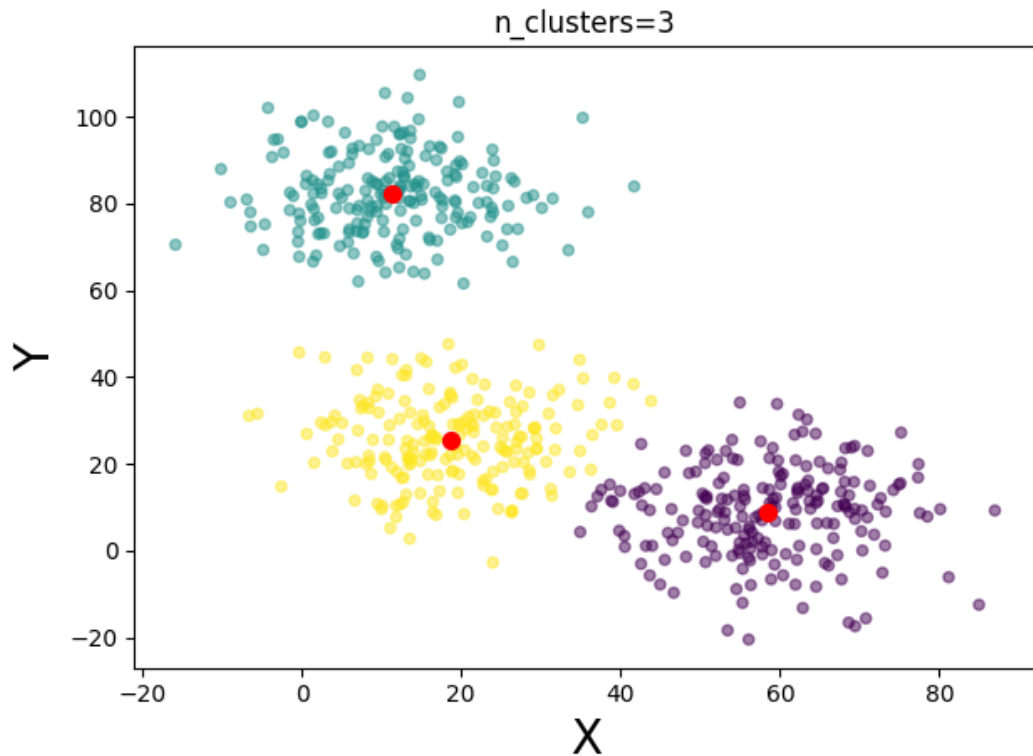
2. Вибірку даних підготував за допомогою генератора випадкових чисел таким чином, щоб утворилось 3 явних кластера даних.

```
my_centers = [[20, 25], [60, 10], [10, 82]]
dots, my_clusters = make_blobs(n_samples=600, centers=my_centers,
cluster_std=10, random_state=0)
df = pd.DataFrame(dots, columns=['X', 'Y'])
```

3. Застосувати кластерний аналіз для підготовлених даних та побудував графік із отриманими кластерами.

```
km = KMeans(n_clusters=3).fit(df)
k_centers = km.cluster_centers_
k_clusters = km.labels_
df['Cluster'] = k_clusters
```

```
plt.scatter(df.X, df.Y, c=k_clusters.astype(float), s=20, alpha=0.5)
plt.scatter(k_centers[:, 0], k_centers[:, 1], c='red', s=50,
alpha=1)
plt.xlabel("X", size=20)
plt.ylabel("Y", size=20)
plt.title("n_clusters=3")
plt.show()
```



4. Провів статистичний аналіз для кожного сформованого кластеру. Як можна замітити з рисунка нижче, кластери були відновлені майже на 100%. З числа точок у кожному кластері, можна зробити припущення, що мінімум 10 точок з 600 не були віднесені до початкових кластерів.

```
clusters_group = df.groupby('Cluster')
print("\nDescribe: \n", clusters_group.describe().T)

plt.scatter(df.X, df.Y, c=k_clusters.astype(float), s=20, alpha=0.5)
plt.scatter(k_centers[:, 0], k_centers[:, 1], c='red', s=50, alpha=1)
```

| Describe: |       |            |            |            |
|-----------|-------|------------|------------|------------|
| Cluster   |       | 0          | 1          | 2          |
| X         | count | 205.000000 | 200.000000 | 195.000000 |
|           | mean  | 58.485649  | 11.442425  | 18.725571  |
|           | std   | 10.112284  | 9.377551   | 9.470811   |
|           | min   | 34.944845  | -15.827966 | -6.591722  |
|           | 25%   | 52.113307  | 5.341054   | 12.455506  |
|           | 50%   | 58.680914  | 11.997772  | 18.640503  |
|           | 75%   | 65.433119  | 17.372421  | 25.775562  |
|           | max   | 86.962241  | 41.709748  | 43.831448  |
| Y         | count | 205.000000 | 200.000000 | 195.000000 |
|           | mean  | 8.793747   | 82.196777  | 25.476847  |
|           | std   | 10.200144  | 9.259805   | 9.664579   |
|           | min   | -20.461431 | 61.526764  | -2.725928  |
|           | 25%   | 1.923515   | 75.999988  | 19.515114  |
|           | 50%   | 9.585992   | 81.714358  | 25.561653  |
|           | 75%   | 14.497121  | 87.458097  | 31.766707  |
|           | max   | 34.124537  | 109.593551 | 47.567235  |

5. Вихідний код програми завантажив у GitHub:  
[https://github.com/DanyloOS/Machine\\_Learning\\_LNU](https://github.com/DanyloOS/Machine_Learning_LNU)

## **Висновок:**

Під час виконання лабораторної роботи я виконав кластерний аналіз та провів статистичний аналіз для кожного сформованого кластеру. Кластерний аналіз може використовуватись в таких сферах, як маркетинг, статистичні дослідження, медицина і т.д.