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

Звіт

з лабораторної роботи №6 з дисципліни
«Аналіз даних в інформаційних системах»

«Класифікація та кластеризація»

Варіант 7

Виконав студент ІП-12 Васильєв Єгор Костянтинович
(шифр, прізвище, ім'я, по батькові)

Перевірив Ліхоузова Тетяна Анатоліївна
(прізвище, ім'я, по батькові)

Київ 2023

Лабораторна робота № 6

Тема: Класифікація та кластеризація

Для виконання лабораторної роботи було обрано мову програмування Python. Спочатку дані було завантажено та імпортовано до датафрейму, а після дослідження, була здійснена їх обробка для подальшої роботи: видалено стовпці PassengerId, Name, Ticket, які не несуть користі для подальших моделей; видалено стовпець Cabin, більшість значень якого є пропущеними; заповнено пропущені значення стовпця Age середнім значенням; заповнено пропущені значення стовпця Embarked модою; переведено категоріальні значення в індикаторні. Після здійснення цих дій датафрейм отримав наступний вигляд:

	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
0	0	3	22.00000	1	0	7.25000	0	1	0	0	1
1	1	1	38.00000	1	0	71.28330	1	0	1	0	0
2	1	3	26.00000	0	0	7.92500	1	0	0	0	1
3	1	1	35.00000	1	0	53.10000	1	0	0	0	1
4	0	3	35.00000	0	0	8.05000	0	1	0	0	1
5	0	3	29.69912	0	0	8.45830	0	1	0	1	0
6	0	1	54.00000	0	0	51.86250	0	1	0	0	1
7	0	3	2.00000	3	1	21.07500	0	1	0	0	1
8	1	3	27.00000	0	2	11.13330	1	0	0	0	1
9	1	2	14.00000	1	0	30.07080	1	0	1	0	0
10	1	3	4.00000	1	1	16.70000	1	0	0	0	1
11	1	1	58.00000	0	0	26.55000	1	0	0	0	1
12	0	3	20.00000	0	0	8.05000	0	1	0	0	1
13	0	3	39.00000	1	5	31.27500	0	1	0	0	1
14	0	3	14.00000	0	0	7.85420	1	0	0	0	1
15	1	2	55.00000	0	0	16.00000	1	0	0	0	1
16	0	3	2.00000	4	1	29.12500	0	1	0	1	0
17	1	2	29.69912	0	0	13.00000	0	1	0	0	1
18	0	3	31.00000	1	0	18.00000	1	0	0	0	1
19	1	3	29.69912	0	0	7.22500	1	0	1	0	0
20	0	2	35.00000	0	0	26.00000	0	1	0	0	1

Рисунок 3.1 – дані у data frame

Було розділено дані на навчальну та тестову вибірки у співвідношенні 80 до 20 за допомогою методу train_test_split бібліотеки sklearn. Було створено й натреновано чотири класифікаційні моделі та порівняно їх за допомогою вбудованого методу score.

```
Logistic regression model accuracy: 86.59%
Decision tree model accuracy: 87.71%
Random forest model accuracy: 87.15%
Gradient boosting model accuracy: 86.59%
```

Рисунок 3.2 – порівняння точності роботи класифікаторів

Метод дерев рішень виявився найкращим, але поліпшити його точність, підбираючи різні гіперпараметри, на жаль, не вдалося.

Для наступного завдання було завантажено нові дані, досліджено структуру даних та знайдено наступні помилки: граматична помилка у назві стовпця Population, числові дані, які частково зберігаються у вигляді рядкових, наявність від'ємних значень та пропущених значень. Після виправлення цих помилок та додавання стовпця з щільністю населення, датафрейм отримав наступний вигляд:

	Country Name	Region	GDP per capita	Population	CO2 emission	Area	Population density
0	Afghanistan	South Asia	561.77875	34656032.00000	9809.22500	652860.00000	53.08341
1	Albania	Europe & Central Asia	4124.98239	2876101.00000	5716.85300	28750.00000	100.03830
2	Algeria	Middle East & North Africa	3916.88157	40606052.00000	145400.21700	2381740.00000	17.04890
3	American Samoa	East Asia & Pacific	11834.74523	55599.00000	165114.11634	200.00000	277.99500
4	Andorra	Europe & Central Asia	36988.62203	77281.00000	462.04200	470.00000	164.42766
5	Angola	Sub-Saharan Africa	3308.70023	28813463.00000	34763.16000	1246700.00000	23.11179
6	Antigua and Barbuda	Latin America & Caribbean	14462.17628	100963.00000	531.71500	440.00000	229.46136
7	Argentina	Latin America & Caribbean	12440.32098	43847430.00000	204024.54600	2780400.00000	15.77019
8	Armenia	Europe & Central Asia	3614.68836	2924816.00000	5529.83600	29740.00000	98.34620
9	Aruba	Latin America & Caribbean	13445.59342	104822.00000	872.74600	180.00000	582.34444
10	Australia	East Asia & Pacific	49755.31548	24127159.00000	361261.83900	7741220.00000	3.11671
11	Austria	Europe & Central Asia	44757.63490	8747358.00000	58712.33700	83879.00000	104.28543
12	Azerbaijan	Europe & Central Asia	3878.70926	9762274.00000	37487.74100	86600.00000	112.72834
13	Bahamas, The	Latin America & Caribbean	28785.47767	391232.00000	2416.55300	13880.00000	28.18674
14	Bahrain	Middle East & North Africa	22579.09342	1425171.00000	31338.18200	771.00000	1848.47082
15	Bangladesh	South Asia	1358.77903	162951560.00000	73189.65300	147630.00000	1103.78351
16	Barbados	Latin America & Caribbean	15891.62655	284996.00000	1272.44900	430.00000	662.78140
17	Belarus	Europe & Central Asia	4989.42776	9507120.00000	63497.77200	207600.00000	45.79538
18	Belgium	Europe & Central Asia	41271.48215	11348159.00000	93350.81900	30530.00000	371.70518
19	Belize	Latin America & Caribbean	4744.73640	366954.00000	495.04500	22970.00000	15.97536
20	Benin	Sub-Saharan Africa	789.44041	10872298.00000	6318.24100	114760.00000	94.73944

Рисунок 3.3 – виправлені дані у data frame

За допомогою «методу ліктя» було візуально визначено оптимальну кількість кластерів (4) для подальшої роботи

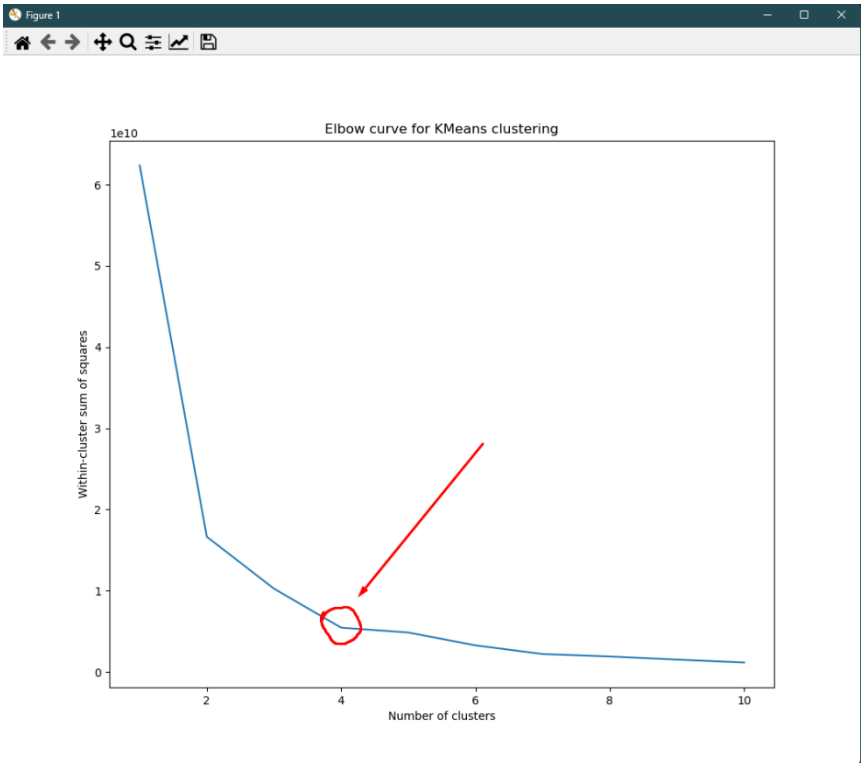

Рисунок 3.4 – візуалізація суми квадратів похибок в залежності від кількості кластерів

Після підбору параметру k для метода KMeans, було розділено параметри GDP per capita та Population density на кластери:

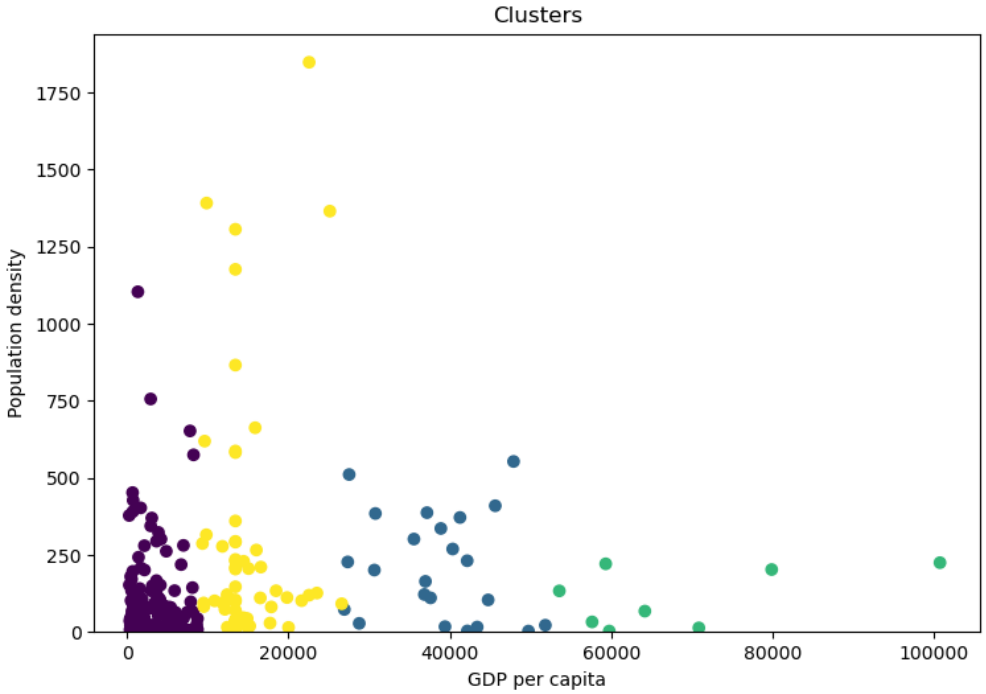

Рисунок 3.5 – візуалізація кластерів

Далі, для кожного кластеру було знайдено домінуючий за кількістю потраплянь регіон:

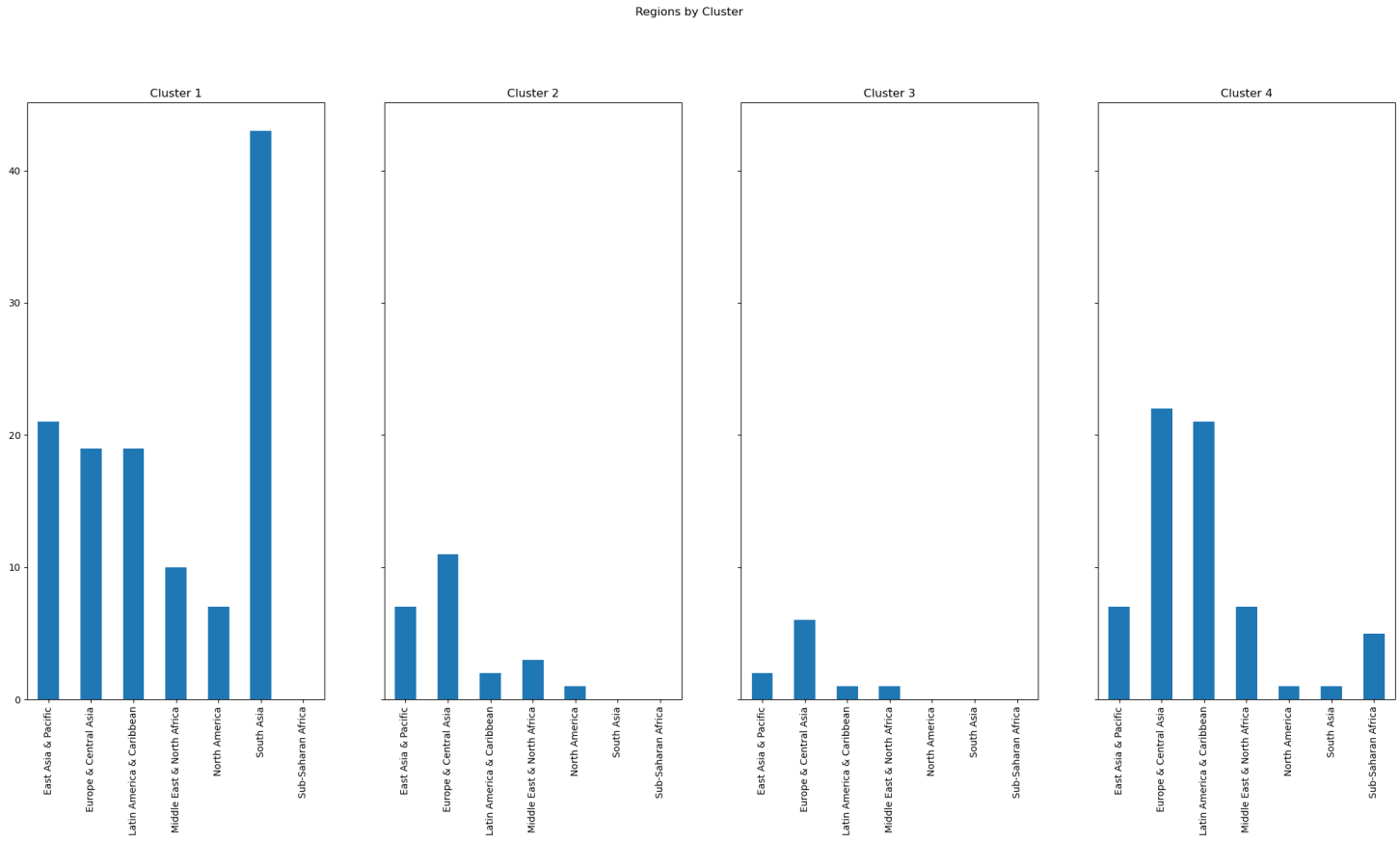

Рисунок 3.6 – кількість входження регіонів до різних кластерів

Також, було визначено загальну кількість потрапляння певного регіону до досліджуваних даних:

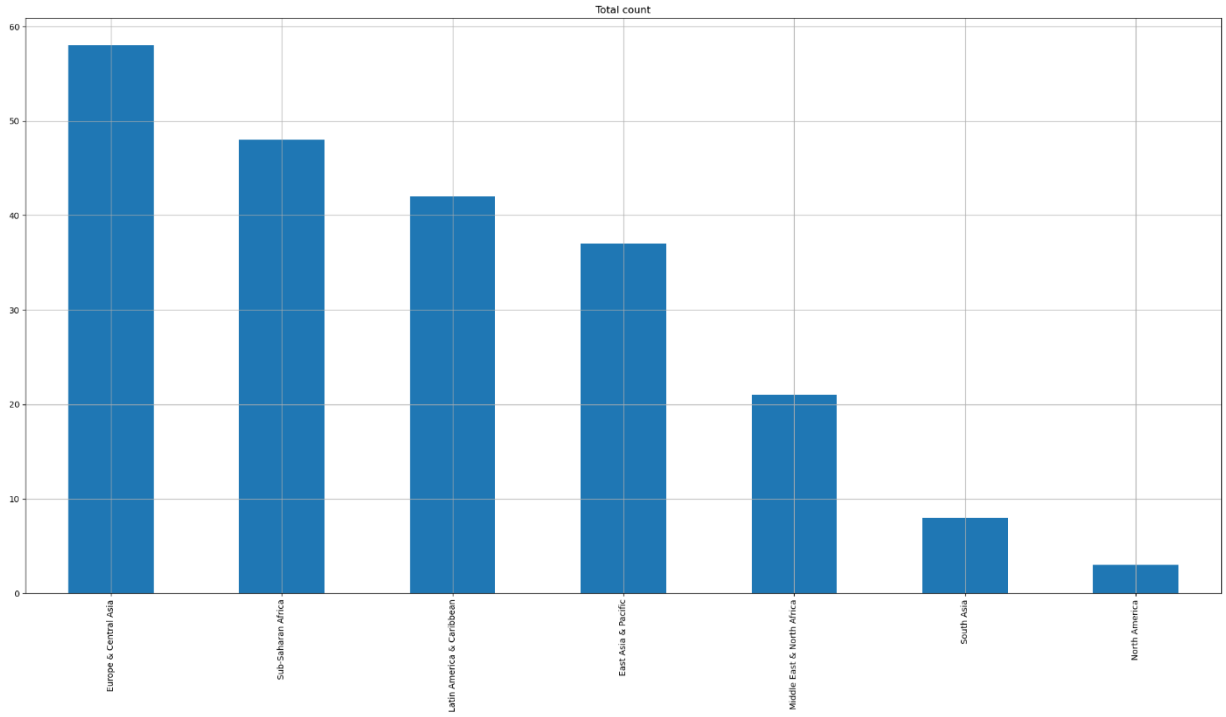

Рисунок 3.7 – стовпчикова діаграма по регіонам

Було побудовано частотні гістограми для кількісних показників у наборі даних, використовуючи цикл:

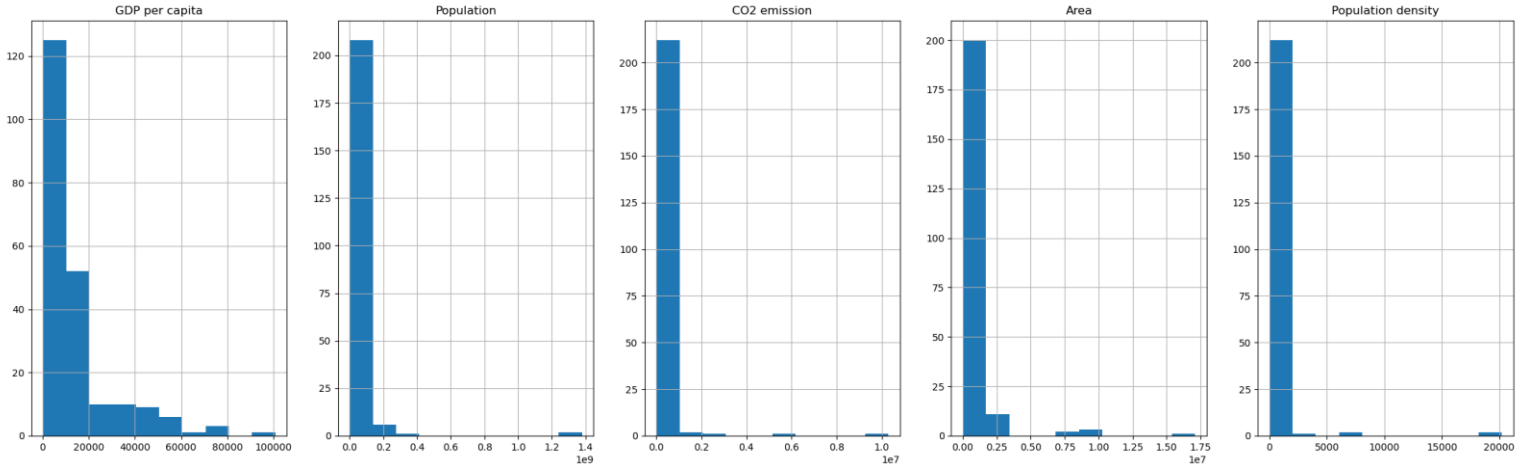

Рисунок 3.8 – візуалізація параметрів data frame

Далі було створено функцію, яка на вхід отримує два набори даних, перевіряє чи є лінійна залежність та повертає відповідно True чи False («лінійна залежність є», якщо коефіцієнт кореляції по модулю більше 0,8)

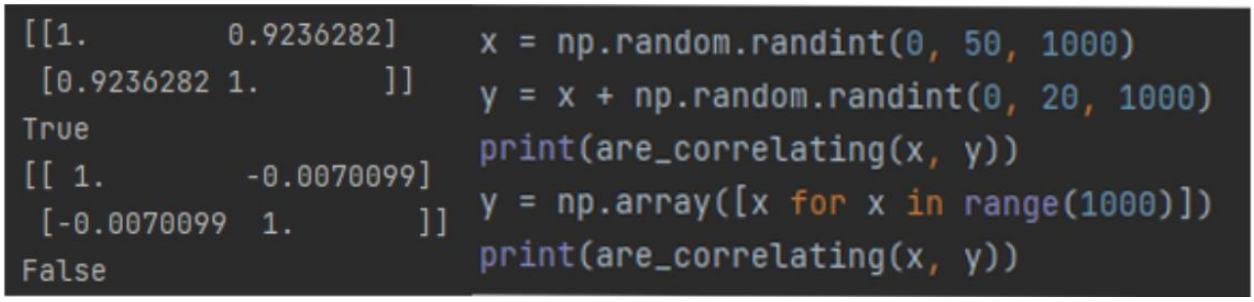

Рисунок 3.9 – результат роботи зазначеної функції

Висновок: досліджуючи данні про вижившних та загиблих пасажирів «Титаніка» було встановлено, що Decision Tree Classifier є найоптимальнішою моделлю для передбачення виживання людини в залежності від інших параметрів, отримана точність 87,71% є доволі непоганою, однак збільшити її не вдалося. Для даних з соціально-економічними показниками країн, 4 кластери стало найкращим параметром для подальшої роботи метода KMeans.